association mapping. Although resolution of the physical position of an allele might be defined by the extent of blocks, assignment of an allelic variant to a particular block should be relatively straightforward. Thus, on a more modest level that possibly excludes fine-mapping, LD association mapping could become more feasible if the notion of blocks is confirmed.

Acknowledgements

I thank D.B. Goldstein for many useful discussions on the implications of blocks for population genetics and G. McVean and H. Nicholas for helpful comments on this manuscript. Financial support from the Wellcome Trust is gratefully acknowledged.

References

1 Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans, Models and data. *Am. J. Hum. Genet.* 69, 1–14

2 Weiss, K.M. and Clark, A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18, 19–24

3 Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144

4 Freimer, N.B. *et al.* (1997) Expanding on population studies. *Nat. Genet.* 17, 371–373

5 Przeworski, M. and Wall, J.D. (2001) Why is there so little intragenic linkage disequilibrium in humans? *Genet. Res.* 77, 143–151

6 Wilson, J.F. and Goldstein, D.B. (2000) Consistent long-range linkage disequilibrium generated by admixture in a Bantu–Semitic hybrid population. *Am. J. Hum. Genet.* 67, 926–935

7 Reich, D.E. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* 411, 199–204

8 Ardlie, K. *et al.* (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.* 69, 582–589

9 Goddard, K.A.B. *et al.* (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* 66, 216–234

10 Jeffreys, A.J. *et al.* (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29, 217–222

11 Rioux, J.D. *et al.* (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* 29, 223–228

12 Daly, M.J. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229–232

13 Patil, N. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294, 1719–1723

14 Gerton, J.L. *et al.* (2000) Inaugural article, global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11383–11390

15 Kirkpatrick, D.T. *et al.* (1999) Maximal stimulation of meiotic recombination by a yeast transcription factor requires the transcription activation domain and a DNA-binding domain. *Genetics* 152, 101–115

16 Johnson, G.C. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29, 233–237

Michael P.H. Stumpf
Dept of Biology, UCL, London,
UK WC1E 6BT.
e-mail: m.stumpf@ucl.ac.uk

Purifying and directional selection in overlapping prokaryotic genes

Igor B. Rogozin, Alexey N. Spiridonov, Alexander V. Sorokin, Yuri I. Wolf, I. King Jordan, Roman L. Tatusov and Eugene V. Koonin

In overlapping genes, the same DNA sequence codes for two proteins using different reading frames. Analysis of overlapping genes can help in understanding the mode of evolution of a coding region from noncoding DNA. We identified 71 pairs of convergent genes, with overlapping 3′ ends longer than 15 nucleotides, that are conserved in at least two prokaryotic genomes. Among the overlap regions, we observed a statistically significant bias towards the 123:132 phase (i.e. the second codon base in one gene facing the degenerate third position in the second gene). This phase ensures the least mutual constraint on nonconservative amino acid replacements in both overlapping coding sequences. The excess of this phase is compatible with directional (positive) selection acting on the overlapping coding regions. This could be a general evolutionary mode for genes emerging from noncoding sequences, in which the protein sequence has not been subject to selection.

DNA sequences can code for more than one gene product by using different reading frames or different initiation codons (Box 1). Overlapping genes are relatively common in DNA and RNA viruses of both prokaryotes and eukaryotes [1–4]. There are several examples in bacterial and eukaryotic genomes, but, in general, overlapping genes are rare other than in viruses [5]. Several studies have addressed the evolution of overlapping genes theoretically and empirically [5–14]. Because of the interdependence of the two overlapping coding regions, the rate of synonymous change appears to be considerably reduced, as is the rate of amino acid changes (nonsynonymous change), although to a lesser extent [8]. Generally, because of the interdependence between the two genes, the rate of mutation fixation is expected to be lower in overlapping regions [7,8,10].

Overlapping genes could evolve as a result of extension of an open reading frame (ORF) caused by a switch to an upstream initiation codon, substitutions in initiation or termination codons, and deletions and frameshifts that eliminate initiation or termination codons [11]. The necessity to maintain two functional overlapping genes inevitably constrains the ability of both genes to become optimally adapted. Such constraints can be alleviated by duplication of the overlapping gene pair, allowing for independent evolution of each gene in the resulting copies. Therefore, overlapping genes can survive long evolutionary spans only when the overlap confers selective advantage to the organism. In viruses, overlapping genes probably persist owing to strong constraints on genome size [5]. In non-viral life forms, the potential advantages of overlapping genes are less clear, although different forms of co-regulation appear to be a possibility [2].

Formation of overlapping genes necessarily involves making a coding region from noncoding DNA. So overlapping genes might help understand *de novo* evolution of coding regions. Which mode of evolution dominates in new coding regions? There seem to be three principal scenarios:

(1) The new protein sequences, in particular the C-terminal regions of overlapping gene products, could be under little functional constraint, evolving neutrally or almost neutrally. Under this model, the overlapping proteins need 'something' at their C-termini to function, the exact sequence is not critical.

(2) A new protein-coding region undergoes directional (positive) selection favoring replacement substitutions, which affect physico-chemical properties of the encoded protein and improve its functional properties (Box 2).

(3) The modes of evolution of the terminal regions of the two overlapping genes might differ; for example, the newly emerging coding sequence could evolve under directional selection, whereas the pre-existing coding sequence in the other partner could be subject to purifying selection.

Analysis of overlapping genes is hampered by sequencing and annotation errors present in genomes [15]. All three types of overlaps between genes (Box 1) can easily emerge because of such errors. Incorrect start codons can lead to 5′-extended ORFs, resulting in artifactual unidirectional or divergent overlaps. Loss of a termination codon caused by a sequencing error can result in an artifactual unidirectional or convergent overlap. Because of this concern, we focused on evolutionarily conserved overlapping gene pairs, which were identified by using the Clusters of Orthologous Groups (COG) database [16] for detecting overlaps that are shared by two or more genomes. However, even among these 'conserved' overlapping genes, a substantial fraction of unidirectional and divergent pairs are likely to be artifacts caused by high rate of mis-annotation of start codons (data not shown). Therefore, all the analysis below deals only with conserved convergent gene overlaps.

A total of 368 conserved, convergent overlapping gene pairs were detected in the analyzed genomes (see supplementary information, ftp://ncbi.nlm.nih.gov/pub/koonin/gene_overlaps/), all of them present in only two species; 127 of these were four-base overlaps that consisted of a stop codon and one coding nucleotide. This type of overlap is common, apparently because the stop codons TAA and TAG provide 'TA' in the complementary chain, which, if completed with an A or a G, also makes a stop codon [11]. Because very short DNA sequences are not amenable to evolutionary analysis, we chose the 71 conserved, overlapping convergent gene pairs with a minimum overlap length of 15 base pairs for further examination, and the 25 conserved gene pairs with overlaps greater than 30 base pairs for more in-depth analyses. Of the 71 analyzed overlaps (see supplementary material), 70 were found in closely related bacterial and archaeal species and only one pair was detected in distantly related genomes (*B. subtilis* – *A. pernix*). Of the 71 overlaps found in two species, 52 were in the same phase (Box 1) in both genomes; in each of these cases, the C-terminal portions

Box 1. Overlapping genes

There are three possible types of adjacent, overlapping genes: unidirectional (the 3′ end of one overlapping with the 5′ end of the other), convergent (the 3′ ends overlapping), and divergent (the 5′ ends overlapping) (Fig. I).

Unidirectional overlapping genes are most widespread, convergent overlapping genes are less common, and divergent overlapping genes are rare. Depending on which codon positions face each other in an overlap, the effects of DNA mutations on the two participating genes can be different. These ways of placing codon positions against each other are termed 'phases' (Fig. II). For each type of overlap, there can be three distinct phases, except for unidirectional overlapping genes, in which only two phases are possible. Our notation for the three possible phases of convergent overlaps is illustrated here. Convergent overlapping genes allow more informative evolutionary analysis because all three phases of overlap have different degrees of dependence between two coding regions, whereas the two possible phases in overlapping unidirectional genes have identical properties [a].

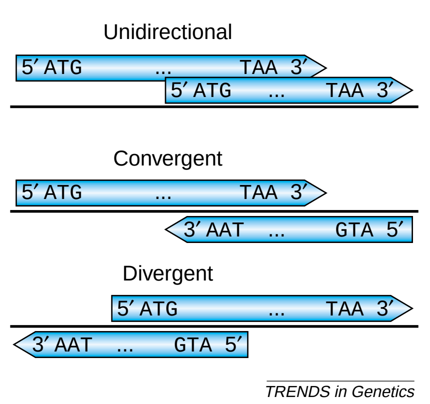


Fig. I. The three classes of overlapping genes.

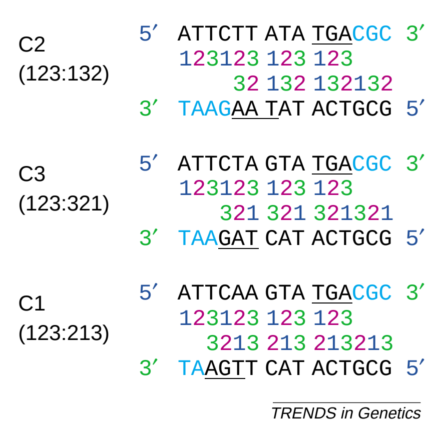


Fig. II. The three phases of convergent overlaps. The numbers denote codon positions. 'C' stands for convergent. Stop codons are underlined.

Reference

a Krakauer, D.C. (2000) Stability and evolution of overlapping genes. *Evolution* 54, 731–739

Box 2. Purifying and directional (positive) selection

Natural selection involves the differential reproductive success of individuals or genotypes in a population. The fitness of a genotype is defined by its ability to reproduce relative to other genotypes in the population. The vast majority of genetic mutations that arise reduce the fitness of the genotypes that bear them. Deleterious alleles produced by mutation are removed from the population by purifying selection. However, a small minority of mutations increases the relative fitness of genotypes. The frequency of the resulting beneficial alleles is increased, and ultimately they are fixed in the population by directional (positive) selection. Thus, purifying selection acts to stabilize allele frequencies, whereas directional selection causes changes in allele frequencies.

Comparisons of protein-coding nucleotide sequences can be used to distinguish between these two types of selection. Such comparisons rely on the analysis of synonymous (S) and nonsynonymous (N) substitution rates. Synonymous changes do not alter the encoded amino acid sequence, whereas nonsynonymous changes result in amino acid replacements. Synonymous changes tend to be (nearly) neutral with respect to fitness and so they are not affected by natural selection. Nonsynonymous changes are most often deleterious and are removed by purifying selection. However, in rare cases, nonsynonymous changes can be beneficial and favored by positive selection. Therefore, the observation of a higher rate of S versus N substitution (S/N > 1) is consistent with purifying selection, whereas a higher relative rate of N substitution (S/N < 1) is consistent with positive selection.

(a)

Thermoplasma acidophilum

```
 D  N  F  S  D  L  V  S  A  A  L  Q  S  Y  E  G  R  Q  D  T  Q  S  L  R  D  R  T  R  R  L  L  Q  R  S  *
GACAACTTCAGCGATCTCGTATCTGCTGCTCTCCAGAGCTATGAAGGTCGTCAAGATACCCAAAGTCTACGAGACCGAACTCGTCGGTTATTGCAAAGATCCTGAAGCAAGA
CTGTTGAAGTCGCTAGAGCATAGACGACGAGAGGTCTCGATACTTCCAGCAGTTCTATGGGTTTCAGATGCTCTGGCTTGAGCAGCCAATAACGTTTCTAGGACTTCGTTCT
        *  R  D  R  I  Q  Q  E  G  S  S  H  L  D  D  L  Y  G  F  D  V  L  G  F  E  D  T  I  A  F  I  R  F  C  S

 S  N  F  N  D  I  V  S  A  A  L  Q  S  Y  E  G  L  R  D  T  Q  S  L  R  D  R  T  R  Q  L  L  Q  K  S  *
AGCAATTTCAACGATATCGTGTCTGCTGCCCTTCAGAGCTACGAAGGTCTTCGAGATACCCAAAGTCTACGAGACCGAACTCGTCAGTTATTGCAAAAATCCTGAAGCTGGA
TCGTTAAAGTTGCTATAGCACAGACGACGGGAAGTCTCGATGCTTCCAGAAGCTCTATGGGTTTCAGATGCTCTGGCTTGAGCAGTCAATAACGTTTTTAGGACTTCGACCT
        *  R  Y  R  T  Q  Q  G  E  S  S  R  L  D  E  L  Y  G  F  D  V  L  G  F  E  D  T  I  A  F  I  R  F  S  S
```

Thermoplasma volcanium

(b)

```
Ta0536_Ta      SDLVSAALQSYEGRQDTQSLRDRTRRLLQRS*
TVN0590_Tv     NDIVSAALQSYEGLRDTQSLRDRTRQLLQKS*
APE2296_Ap     REAVELALNSY-----TKKVGGALRRLLEEA...
Consensus100%  p-hh.hALpSY.....Tpph....RpLLpps
```

(c)

```
Ta0537m_Ta     FRI-FAITDEFGLVDFGYLDDLHSSGEQQIRDR*
TVN0591_Tv     FRI-FAITDEFGLVDFGYLEDLRSSEGQQTRYR*
APE1861_Ap     EAVPIIIDESVGSLKLPPLRELDKILG*
AF1155_Af      IEV-YAIEEE—-VVYLGSIEDLRKII*
MTH1803_Mth    FLV-YGVEDF-EIFEFGTVGDLLKFQQKTGYPGD*
MJ0037_Mj      AEV-IAITDI-GLLNFGTLRDLREFAKTHL*
PH1310_Ph      LEV-LVTTGD-ELLNFGRFSQLIE-AMKRL*
PAB0562_Pab    LSV-VATTGE-ELLNFGKFGDLIR-AMRLLG*
Consensus80%   h.h.hh.p....hhphG.h.-L.p.......
```

TRENDS in Genetics

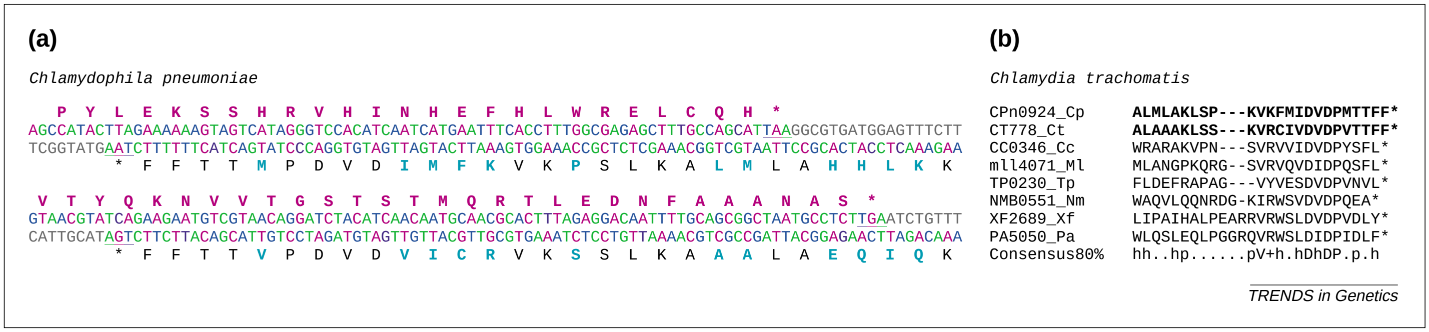
Fig. 1. Overlap between COG0373 and COG1407 genes in *Thermoplasma acidophilum* and *Thermoplasma volcanium*. (a) Arrangement of overlapping regions and amino acid conservation within the overlap between the two *Thermoplasma* species. Amino acid differences are shown in magenta for COG0373 proteins and in cyan for COG1407 proteins. (b) Multiple alignment of the C-termini of COG0373 proteins. The alignment is a portion of a complete alignment of the corresponding proteins, which was constructed using the T_Coffee program [18]. The sequences from the overlapping region in *Thermoplasma* are shown in bold. The consensus shows: h, hydrophobic residues (ACVILMFYW); p, polar residues (STDENQKRH); s, small residues (GASDNCV); and –, negatively charged residues (DE). A dot shows no consensus for the given position. The proteins are designated by the systematic gene name and species abbreviation: Ta, *Thermoplasma acidophilum*; Tv, *Thermoplasma volcanii*; Ap, *Aeropyrum pernix*. The alignment in the C-terminal block shown in the figure was statistically significant ($P<10^{-4}$) as determined by using the MACAW program [19]. (c) Multiple alignment of the C-termini of COG1407 proteins. The alignment in the C-terminal block shown in the figure was statistically significant ($P<10^{-11}$) as determined by using the MACAW program. The designations are as in (b). Additional species: Af, *Archaeoglobus fulgidus*; Mj, *Methanococcus jannaschii*; Mth, *Methanobacterium thermoautotrophicum*; Ph, *Pyrococcus horikoshii*; Pab, *Pyrococcus abyssi.*

of both proteins in question encoded within the overlap showed sequence conservation continuous with the conservation in the rest of the protein (data not shown), which suggests that they emerged in the common ancestor of the respective species. Some of the remaining overlaps, in particular, the one in *B. subtilis* and *A. pernix*, might, in principle, have evolved independently between pairs of genes from the same two COGs.

An example of a conserved convergent overlap in phase C2 (123:132) (Box 1) involves glutamyl-tRNA reductase (COG0373) and ICC-like phosphoesterase (COG1407) genes from two *Thermoplasma* species. In this case, both species have an overlap of 98 base pairs (Fig. 1a). The C-termini of both proteins show statistically significant sequence conservation with the corresponding orthologous genes from other archaeal species (Fig. 1b), making it impossible to determine which gene was extended to create the overlap. Given the sequence conservation in the overlapping regions, it seems probable that this overlap emerged early during archaeal evolution and still persists in the *Thermoplasma* species, whereas, in the other sequenced archaeal genomes, the gene pair was disjointed, perhaps by gene duplication. It should be noted, however, that this was the only example of long-range evolutionary conservation of both C-terminal protein regions within overlaps, which suggests that most of the overlaps evolved relatively recently.

A convergent overlap between the genes for 7-keto-8-aminopelargonate synthetase (COG0156) and a Superfamily II helicase PriA (COG1198) is conserved in six species of the family *Chlamydiaceae* (Fig. 2a). The overlap is in phase C3 (123:321) in *Chlamydophila pneumoniae* and *Chlamydophila psittaci* and in phase C1 (123:213) (Box 1) in *Chlamydia trachomatis* and *Chlamydia muridarum*. The overlapping carboxyl-ends of COG1198 proteins have regions of statistically significant conservation with some orthologous bacterial helicases (Fig. 2b). By contrast, no sequence conservation was found for the overlapping C-ends of COG0156 proteins even between the chlamydial species themselves (Fig. 2a and data not shown). Thus, the overlap might have evolved by extension of a COG0156 gene into the coding region of the COG1198 gene because of the loss of the stop codon in the common ancestor of the chlamydial species, with a subsequent frameshift in one of the lineages. A less likely alternative is the independent origin of the overlap between the same pair of orthologous genes in the common ancestors of *C. pneumoniae* and *C. psittaci*, on the one hand, and of *C. trachomatis* and *C. muridarum*, on the other hand.

We conducted a test for purifying selection to assess the functional importance of the overlapping regions for the corresponding genes. Standard methods for analysis of modes of natural selection are based on the synonymous–nonsynonymous substitution ratio in coding regions (Box 2). However, methods that are normally used for estimating the substitution rates are not applicable to overlaps between genes, because of their interdependence; in other words, the nucleotide substitutions in overlapping regions can be considered synonymous or nonsynonymous only with regard to a particular reading frame [17]. To avoid this dependence, we analyzed fourfold degenerate third positions of codons in phase C2 (123:132). In this phase, second codon positions, in which all mutations lead to amino acid replacements, are located opposite third, degenerate codon positions of a complementary coding region. We assumed that purifying selection in fourfold degenerate sites is very


(a)
Chlamydophila pneumoniae
P Y L E K S S H R V H I N H E F H L W R E L C Q H *
AGCCATACTTAGAAAAAAGTAGTCATAGGGTCCACATCAATCATGAATTTCACCTTTGGCGAGAGCTTTGCCAGCATTAAGGCGTGATGGAGTTTCTT
TCGGTATGAATCTTTTTTCATCAGTATCCCAGGTGTAGTTAGTACTTAAAGTGGAAACCGCTCTCGAAACGGTCGTAATTCCGCACTACCTCAAAGAA
* F F T T M P D V D I M F K V K P S L K A L M L A H H L K K
V T Y Q K N V V T G S T S T M Q R T L E D N F A A A N A S *
GTAACGTATCAGAAGAATGTCGTAACAGGATCTACATCAACAATGCAACGCACTTTAGAGGACAATTTTGCAGCGGCTAATGCCTCTTGAATCTGTTT
CATTGCATAGTCTTCTTACAGCATTGTCCTAGATGTAGTTGTTACGTTGCGTGAAATCTCCTGTTAAAACGTCGCCGATTACGGAGAACTTAGACAAA
* F F T T V P D V D V I C R V K S S L K A A A L A E Q I Q K
(b)
Chlamydia trachomatis
CPn0924_Cp ALMLAKLSP---KVKFMIDVDPMTTFF*
CT778_Ct ALAAAKLSS---KVRCIVDVDPVTTFF*
CC0346_Cc WRARAKVPN---SVRVVIDVDPYSFL*
mll4071_Ml MLANGPKQRG--SVRVQVDIDPQSFL*
TP0230_Tp FLDEFRAPAG---VYVESDVDPVNVL*
NMB0551_Nm WAQVLQQNRDG-KIRWSVDVDPQEA*
XF2689_Xf LIPAIHALPEARRVRWSLDVDPVDLY*
PA5050_Pa WLQSLEQLPGGRQVRWSLDIDPIDLF*
Consensus80% hh..hp......pV+h.hDhDP.p.h

TRENDS in Genetics

Fig. 2. Overlap between COG0156 and COG1198 genes in *Chlamydophila pneumoniae* and *Chlamydia trachomatis*. (a) Arrangement of overlapping regions and amino acid conservation within the overlap between the two *Chlamydia* species. Amino acid differences are shown for COG1198 proteins are shown in cyan; the entire sequences of the COG0156 proteins are shown in magenta to emphasize the lack of sequence conservation within the overlap. (b) Multiple alignment of the C-termini of COG1198 proteins (the PriA helicase). The alignment in the C-terminal block shown in the figure was statistically significant ($P < 10^{-18}$) as determined by using the MACAW program. The designations are as in Fig. 1b,c. Species abbreviations: Cp, *Chlamydophila pneumoniae*; Ct, *Chlamydia trachomatis*; Cc, *Caulobacter crescentus*; Ml, *Mesorhizobium loti*; Tp, *Treponema pallidum*; Nm, *Neisseria meningitides*; Xf, *Xylella fastidiosum*; Pa, *Pseudomonas aeruginosa*.

weak (or absent) and acts equally in overlapping and non-overlapping coding regions. Then, any difference between the rates of mutations in such silent positions between overlapping and non-overlapping parts of a gene should reflect selection acting on the second codon positions of the complementary coding region. In individual overlaps, the number of fourfold degenerate sites was small, except for the pair COG0373–COG1407 (Fig. 1). In this case, the fraction of different nucleotides between the two *Thermoplasma* species for COG0373 in fourfold sites was 0.21 (3/14) and 0.60 (59/98) in overlapping and non-overlapping regions, respectively. For COG1407, the corresponding values were 0.27 (3/11) and 0.71 (46/65), respectively. Altogether, fourfold degenerate sites in 15 overlap regions of eight overlapping genes (in one gene, no fourfold degenerate sites were found) with overlap ≥30 base pairs in both species had a significantly lower nucleotide substitution rate than the fourfold degenerate sites in the non-overlapping regions of the same genes ($P < 0.001$ under the binomial distribution). Thus, the regions of both genes within an overlap appear to be subject to purifying selection, indicating that the protein sequences encoded in the overlapping regions are functionally important.

Notably, over a half of the overlaps were found in phase C2 (123:132) (Table 1). We estimated the probability of such an excess of phase C2 under the assumption that the three phases are equally likely and using a more realistic model, which accounted for the expected lengths of potential extensions of a convergent coding region into the coding region of the given gene in different phases (distance to the next stop). The lengths of potential extensions were determined for all members of the COGs that form the 25 overlaps ≥30 bp. A decreased length of extensions in phase C1 (123:213) was observed, probably owing to the fact that the complementary sequences of the stop codons TAG/TAA always include the TA dinucleotide, which can also be a part of the stop codon in the complementary strand [11]. The distributions of potential extension lengths in phases C2 (123:132) and C3 (123:321) were approximately the same (data not shown). The numbers of potential extensions in different phases that exceeded a given length cut-off were used to calculate the expected phase distribution for the conserved convergent overlaps (Table 1).

In each case, statistical analysis using the χ^2 test showed that the deviation of the observed phase distribution (manifest primarily in the excess of the C2 phase) from the expected distribution was highly statistically significant for overlaps ≥15 bp (Table 1). There is no reason to believe that the processes, through which overlaps emerge, would favor one phase over another. By contrast, selective pressure might affect the distribution of the phases because an overlap restricts the potential for adaptation attainable for the given protein through changing amino acids in specific sites. The three phases vary in how strongly coupled are the amino acid sequences encoded in the two overlapping genes within the overlap. The C2 (123:132) phase permits the most amino acid replacements in one gene's overlapping region, without affecting the amino acid sequence of the other gene's overlapping region. The C3 phase (123:321) is more constrained, and C1 (123:213) allows fewer non-disruptive amino acid changes than the other two phases [10].

Because in phase C2 (123:132), the second codon position in one coding sequence faces the third (degenerate) position in the other sequence, one should expect that the majority of amino acid replacements within the overlapping regions are brought about by mutations in the second positions. We tested this prediction by comparing the contribution of mutations in different codon positions to amino acid replacements in the overlaps

Table 1. Phase distributions for observed and potential convergent overlaps[a]

Observed[b]				χ^2	P[c]	Expected[d]			
Minimal overlap length (bp)	Phase C1 123:213	 C2 123:132	 C3 123:321			Phase C1 123:213	 C2 123:132	 C3 123:321	Minimal overlap length (bp)
15	17	71	31	31.0	1.9×10^{-7}	32.9	42.3	43.8	15
30	8	22	11	7.2	0.027	11.9	13.9	15.2	30
45	8	13	8	2.4	0.150	8.7	9.3	11.0	45
60	3	9	6	3.8	0.247	5.6	5.4	7.0	60
75	3	4	4	0.4	0.819	3.6	3.0	4.4	75
90	0	4	2	3.7	0.058	2.0	1.6	2.4	90

[a] All conserved convergent overlaps longer than 15 bp were analyzed.
[b] The distribution of the phases of conserved convergent overlaps.
[c] Only the value for overlaps ≥15 bp should be considered significant if multiple comparisons are taken into account.
[d] Distribution of phases for potential convergent extensions in all clusters of orthologous genes (COGs) analyzed in this study normalized for the number of conserved convergent overlaps.

Table 2. Amino acid replacements resulting from mutations in different codon positions in overlapping and non-overlapping regions

	Number of replacements (% of the total)[a]		
	First position	Second position	Third position
Overlaps	9 (19)	33 (69)	6 (12)
Non-overlapping regions	194 (45)	134 (31)	103 (24)

[a]The contribution of replacements in the second codon position to the total number of replacements was significantly greater (P<0.001; Fisher exact test) for the overlapping than for the non-overlapping regions. Conversely, the contribution of replacements in the first codon position was significantly greater for the non-overlapping regions.

and in the non-overlapping regions of the same genes. As predicted, in the C2-phase overlaps, the second position contributed the most, in contrast to the non-overlapping regions where the majority of nonsynonymous substitutions occurred in the first position (this reflects purifying selection, which tends to eliminate nonconservative amino acid replacements resulting from nucleotide substitutions in the second position) (Table 2). Thus, the substitution pattern in the overlapping regions is compatible with the notion that coupled amino acid replacements are strongly selected against.

The observed prevalence of phase C2 (123:132) is an indication that it is evolutionarily advantageous to allow nonconservative (and uncoupled) amino acid replacements in both genes in the overlapping region. There seems to be a good justification for allowing nonconservative changes within the overlaps: they can produce advantageous mutations in newly formed coding regions (when an overlap emerges, the protein sequence encoded by a previously noncoding region is unlikely to be functionally adapted). In other words, it appears that, at least at early stages of the evolution of new protein sequences within overlaps, directional (positive) selection is a major factor. The evolutionary scenario suggested by these observations, with positive selection acting early after the emergence of a new protein sequence and purifying selection taking over at subsequent stages of evolution, might have a general impact on our understanding of *de novo* evolution of proteins.

Supplementary information

A complete list of analyzed prokaryotic genomes, a complete list of conserved convergent overlaps, and an annotated table summarizing the properties of overlaps ≥30 bp are available at ftp://ncbi.nlm.nih.gov/pub/koonin/gene_overlaps/.

Acknowledgements

We thank Darren Natale and Fyodor Kondrashov for helpful discussions.

References

1 Barrell, B.G. *et al.* (1976) Overlapping genes in bacteriophage phiX174. *Nature* 264, 34–41

2 Normark, S. *et al.* (1983) Overlapping genes. *Annu. Rev. Genet.* 17, 499–525

3 Lamb, R.A. and Horvath, C.M. (1991) Diversity of coding strategies in influenza viruses. *Trends Genet.* 7, 261–266

4 Samuel, C.E. (1989) Polycistronic animal virus mRNAs. *Prog. Nucleic Acid Res. Mol. Biol.* 37, 127–153

5 Keese, P.K. and Gibbs, A. (1992) Origins of genes: 'big bang' or continuous creation? *Proc. Natl. Acad. Sci. U. S. A.* 89, 9489–9493

6 Sander, C. and Schulz, G.E. (1979) Degeneracy of the information contained in amino acid sequences: evidence from overlaid genes. *J. Mol. Evol.* 13, 245–252

7 Pavesi, A. *et al.* (1997) On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. *J. Mol. Evol.* 44, 625–631

8 Miyata, T. and Yasunaga, T. (1978) Evolution of overlapping genes. *Nature* 272, 532–535

9 Kozlov, N.N. (1999) Demand of each of 64 codons in genetic overlapping areas. *Dokl. Akad. Nauk* 367, 544–547

10 Krakauer, D.C. (2000) Stability and evolution of overlapping genes. *Evolution* 54, 731–739

11 Fukuda, Y. *et al.* (1999) Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 27, 1847–1853

12 Jordan, I.K. *et al.* (2000) Molecular evolution of the Paramyxoviridae and Rhabdoviridae multiple-protein-encoding P gene. *Mol. Biol. Evol.* 17, 75–86

13 Smith, T.F. and Waterman, M.S. (1981) Overlapping genes and information theory. *J. Theor. Biol.* 91, 379–380

14 Shcherbakov, D.V. and Garber, M.B. (2000) Overlapping genes in bacterial and bacteriophage genomes. *Mol. Biol. (Mosk.)* 34, 572–583

15 Natale, D.A. *et al.* (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica* 108, 9–17

16 Tatusov, R.L. *et al.* (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36

17 Hein, J. and Stovlbaek, J. (1995) A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. *J. Mol. Evol.* 40, 181–189

18 Notredame, C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217

19 Schuler, G.D. *et al.* (1991) A workbench for multiple alignment construction and analysis. *Proteins* 9, 180–190

Igor B. Rogozin
Alexey N. Spiridonov
Alexander V. Sorokin
Yuri I. Wolf
I. King Jordan
Roman L. Tatusov
Eugene V. Koonin*
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
*e-mail: koonin@ncbi.nlm.nih.gov

Growth and decline of introns

Alexander E. Vinogradov

To gauge the processes that might direct the length of introns, I studied the balance of indels (insertions or deletions, determined using Alu and LINE1 retroposon repeats) and the density of these repeats in the introns of the human genome. The indel balance is biased in favour of deletions and correlated with the divergence of repeats. At fixed repeat divergence, the indel bias correlated with the intron size: the shorter the intron, the more deletions were favoured over insertions. This correlation with the intron size was stronger than with the gene-wide or isochore-wide parameters. The density of repeats (the number of repeats in a unit of intron length) correlated positively with the intron size. Thus, quite different mechanisms, the indel bias and the integration and/or persistence of retroposons, act in the same direction in regards to intron size, which suggests selection for the size of individual introns.

0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(02)02660-4