Origin and evolution of overlapping genes in the family Microviridae

Angelo Pavesi

Department of Genetics, Anthropology and Evolution, University of Parma, Parco Area delle Scienze 11/A, I-43100 Parma, Italy

The possibility of creating novel genes from pre-existing sequences, known as overprinting, is a widespread phenomenon in small viruses. Here, the origin and evolution of gene overlap in the bacteriophages belonging to the family Microviridae have been investigated. The distinction between ancestral and derived frames was carried out by comparing the patterns of codon usage in overlapping and non-overlapping genes. By this approach, a gradual increase in complexity of the phage genome – from an ancestral state lacking gene overlap to a derived state with a high density of genetic information – was inferred. Genes encoding less-essential proteins, yet playing a role in phage growth and diffusion, were predicted to be novel genes that originated by overprinting. Evaluation of the rates of synonymous and non-synonymous substitution yielded evidence for overlapping genes under positive selection in one frame and purifying selection in the alternative frame.

In viruses under strong pressure to minimize genome size, overlapping genes represent a fine strategy to condense a maximum amount of information into short nucleotide sequences (reviewed by Gibbs & Keese, 1995). The possibility of creating novel genes from pre-existing sequences, known as overprinting, has long attracted the attention of researchers (Miyata & Yasunaga, 1978; Sander & Schulz, 1979). General properties of gene overlap have been investigated with information theory (Pavesi et al., 1997) and mathematical models (Krakauer, 2000).

It has been claimed that strong constraints should be the rule in overlapping genes, as a single mutation will impair two amino acid sequences. A typical example of constrained evolution concerns the overlapping envelope and polymerase genes of Hepatitis B virus (Mizokami et al., 1997). The detection of unusually strong constraints at the third codon position in the large open reading frame (ORF) of Hepatitis C virus and Hepatitis G virus has supported the hypothesis that an additional protein is encoded by the overlapping frame (Pavesi, 2000; Walewski et al., 2001). The same feature has been associated, in gene P of Vesicular stomatitis virus, with the existence of a new overlapping ORF (Spiropoulou & Nichol, 1993).

Some recent papers, however, indicate that overlapping genes can exhibit a more flexible pattern of change. Simian immunodeficiency virus (Hughes et al., 2001), Potato leafroll virus (Guyader & Ducray, 2002) and Human papillomavirus (Narechania et al., 2005) all show a high rate of non-synonymous change in one reading frame (positive selection) with concurrent dominance of synonymous substitutions in the alternative frame (purifying selection).

Although several cases of gene overlap have been identified, less effort has been devoted to determining their origins. Usually, this task is performed with a phylogenetic approach, which compares homologous genes from a wide number of virus families. The presence of gene overlap in a given family, and its lack in all others, supports the following hypothesis: the ancestral frame is that shared by all viruses, whilst the frame that originated by overprinting is typical of a few closely related viruses. By this approach, the origin of overlapping genes has been clarified in tymoviruses, luteoviruses, lentiviruses and paramyxoviruses (Keese & Gibbs, 1992; Jordan et al., 2000).

In this paper, we have investigated the origin and evolution of gene overlap in the bacteriophages that belong to the family Microviridae and infect Escherichia coli. The coliphages ΦX174, z3 and G4 show a similar genome structure, and the most convincing evidence of homology comes from the genes that overlap. In all phages, gene E is encoded entirely within gene D; likewise, gene B lies within gene A and gene K lies within genes A and C (Sanger et al., 1977; Godson et al., 1978; Kodaira et al., 1992).

Following the suggestion that out-of-frame expression of a gene often entails a bias at the third codon position (Keese & Gibbs, 1992), the origin of gene overlap was investigated by comparing the patterns of codon usage in overlapping and non-overlapping genes. This approach was preferred to the phylogenetic method described above, because the microviruses that infect E. coli present a genome structure rather different from those that parasitize Chlamydia psittaci, Spiroplasma melliferum and Bdellovibrio bacteriovorus (Renaudin et al., 1987; Storey et al., 1989; Brentlinger et al., 2002).
In total, 30 complete genome sequences of coliphages were collected from GenBank (accession nos AF176027–AF176034, AF299300–AF299314, J02482, AF274751, M14428, AF454431, V00657 and X60322–X60323). This corpus of data contains 24 isolates from ΦX174, two isolates from S13, two isolates from G4 and one isolate from each of W3 and ΦK, respectively.

The protein-coding region of each genome sequence was separated into 13 segments. Four of them correspond to the non-overlapping genes J, F, G and H. Genes A, C and D were subdivided into the corresponding overlapping and non-overlapping regions, yielding a total of six segments. Genes B, E and K, all embedded entirely within other genes, yielded the remaining three segments.

The use of synonymous codons was evaluated with the relative synonymous-codon usage (RSCU) index (Sharp & Li, 1987). For each of the 59 degenerate codons, the RSCU value was calculated as follows:

\[
\text{RSCU} = \left( \frac{N_{\text{codon}}}{N_{\text{amino~acid}}} \right) D
\]

where \(N_{\text{codon}}\) is the frequency of a codon in a given coding sequence, \(N_{\text{amino~acid}}\) is the frequency of the amino acid specified by that codon and its synonyms and \(D\) is the degeneracy of that amino acid (e.g. the degeneracy value is 6 for Arg, 3 for Ile and 2 for Tyr). The RSCU data were included into a matrix of 13 rows (the number of coding segments) and 59 columns (the number of codons). This matrix was subjected to principal-component analysis (PCA).

PCA is a multivariate statistical method for simplifying the multi-dimensional information of the data matrix into a two-dimensional map (Morrison, 1976). This procedure guarantees that the highest principal components, usually the first and the second, contain as much information as possible. Further components will recover smaller and smaller amounts of the residual variation. Given the eigenvalues of the data matrix, the percentages of information recovered by the various components can be calculated.

The main features of the pattern of codon usage in coliphages are illustrated in the two-dimensional map yielded by PCA (Fig. 1). The projection of points on axis 1, which accounts for 30 % of the total information, separated the 13 coding regions into two groups. Such a clustering reflects two distinct patterns of codon usage. Points at the extreme left of axis 1 (position-coordinate values from \(-3.1\) to \(-5.5\)) correspond to the non-overlapping genes H, F and J. Points at the extreme right of axis 1 (position-coordinate values of \(5.2\) and \(5.8\)) correspond to the overlapping genes E and K.

Following the hypothesis that the use of synonyms in non-overlapping genes reflects the ancestral pattern of codon usage, genes E and K should be viewed as novel genes that arose later by overprinting. A de novo origin of gene E is also supported by the position on axis 1 of the overlapping region of gene D (\(-2.3\)). Such a position highlights the fitting of gene D to the original pattern of codon usage. Ancestry of the entire gene D is confirmed by the close proximity between its overlapping and non-overlapping

**Fig. 1.** PCA map of the pattern of codon usage in coliphages. Axis 1 accounts for 30 % and axis 2 for 16 % of the total variation in the data matrix. The subsequent axes, accounting for a progressively smaller amount of the residual variation, did not provide more relevant information on the pattern of codon usage (data not shown). Abbreviations indicate overlap (ov.) and non-overlap (non-ov.).
parts (position-coordinate values of −2·3 and −1·9, respectively).

The map positions of genes A and B supported a different evolutionary pathway. Unlike gene D, the overlapping region of gene A was placed at a distance from its non-overlapping counterpart (position-coordinate values of 3·3 and −1·4, respectively). This result proves the existence in gene A of two distinct patterns of codon usage, one similar to ancestral genes and the other to derived genes. Gene B shows a use of synonyms roughly similar to that of non-overlapping genes, because it was placed at the beginning of the left half of axis 1 (position-coordinate value of −0·52). These findings suggest that the coliphage genome originally possessed a short gene A, located in close proximity to gene B. Gene A probably reached its present length by taking advantage of a new termination codon beyond gene B.

Like gene A, the overlapping and non-overlapping regions of gene C were placed at a distance from each other (position-coordinate values of 4·1 and 0·6, respectively). Again, an original shorter gene C, which evolved by using as initiator codon an upstream ATG localized at the end of gene A, can be hypothesized.

The projection of points on axis 2, which accounts for 16% of the total information, separated the overlapping gene K (position-coordinate value of −6·6) from all of the other coding regions (position-coordinate values from 4·0 to −1·9). The distinctiveness of gene K at both the first and second axes of ordination reflects a use of synonyms that is very different from that of all other genes, thus further supporting the view of a de novo origin by overprinting. The gradual increase in complexity of the phage genome—from a putative ancestral state lacking gene overlap to a derived state with a high density of genetic information—is summarized in Fig. 2.

The type of selective pressure affecting the evolution of overlapping genes was investigated by evaluating the occurrence of both synonymous and non-synonymous substitutions. The best candidates for this analysis were genes A/B and D/E. Gene A, for example, contains both a large non-overlapping region and a shorter region that overlaps the entire gene B. This feature allows us to test the hypothesis that different selective pressures act on the same gene.

Multiple alignment of the sequences of genes A/B and D/E was carried out with the CLUSTAL W program (Thompson et al., 1994). Sequences were taken from the most divergent coliphages (ΦX174, α3 and G4), after exclusion of redundant data. The predicted amino acid sequences were first aligned in CLUSTAL W and the respective codons were then placed on them. The amino-terminal region of protein A, at which the alignment postulated large insertions or deletions, was excluded. The rates of synonymous and non-synonymous substitution were estimated by the method of Nei & Gojobori (1986).

As reported in Table 1, the non-overlapping region of gene A shows a rate of non-synonymous substitution per site (Ks) about six times lower than that of synonymous substitution (Kn), yielding a Ks/Kn ratio significantly lower than unity (0·16). As a similar Ks/Kn ratio (0·33) was also found in the region of overlap, it can be hypothesized that a substantial proportion of amino acid changes in protein A must have been eliminated by purifying selection. The rate of non-synonymous change in gene B (Ks = 0·56) is due to the overlap between the third codon position of gene A and the first codon position of gene B. Being greater than unity, the Ks/Kn ratio found in gene B (1·24) is indicative of adaptive or positive selection.

Analysis of genes D and E revealed a similar evolutionary pattern. Both the overlapping and non-overlapping regions of gene D exhibited a rate of non-synonymous change lower than that of synonymous change. The corresponding Ks/Kn ratios were both considerably lower than unity (0·16 and 0·18, respectively), thus suggesting purifying selection. Adaptive selection was hypothesized for gene E, because of a Kn value (0·27) higher than Kn (0·19). Synonymous substitutions at the third codon position of frame D, in fact,
produce non-synonymous changes at the second codon position of frame E.

These findings suggest that the rate of nucleotide change in the two-phase coding regions of genes A and D, albeit considerably lower than that of the one-phase coding regions (see the absolute values of $K_a$ and $K_s$ in Table 1), does not preclude a pattern of adaptive evolution in the corresponding overlapping genes, B and E.

Finally, the origins of genes A/B and D/E were further investigated by using Pearson’s correlation test, with the aim of validating the results provided by PCA. The RSCU values from each overlapping gene were compared with those obtained from the whole set of non-overlapping genes. A highly significant correlation ($P < 0.001$) was found in genes B and D ($r$ values of 0.53 and 0.69, respectively), whilst a lack of correlation was found in gene E ($r = 0.01$). These findings confirm the hypothesis that genes B and D are ancestral genes, whereas gene E is a more-recent gene.

Analysis of gene A yielded two remarkably different $r$ values: a high degree of correlation in the non-overlapping part ($r = 0.88$) and a poor degree of correlation in the overlapping part ($r = 0.02$). This result is in accordance with the hypothesis stated above, that is, a shorter gene A that evolved by using the sequence of gene B in a different translational frame.

The findings on overlapping genes presented here can be discussed by taking into account their role during the infectious cycle. For example, the overlapping gene E plays a crucial role in the final step of infection, as it encodes a protein causing lysis of the host cell (Bläsi & Lubitz, 1985). It is important to note that protein E is not an essential structural component of the phage, as normal phage particles are produced in the absence of lysis (Hutchinson & Sinsheimer, 1966). Thus, the acquisition of gene E by overprinting can be viewed as an evolutionary advantage favouring the diffusion of mature phages. Interestingly, a similar role has been assigned to a new overlapping gene found in tymoviruses (Bozarth et al., 1992).

Another gene that arose by overprinting, gene K, also encodes a less-essential protein, as demonstrated by the finding that mutants of φX174 are viable even when they make no detectable K protein (Tessman et al., 1980). However, a study by Gillam et al. (1985) assigned a phenotype to protein K, as it demonstrated that mutant phages lacking gene K show a burst size sixfold lower than that of wild-type phages. Again, a beneficial effect for phage growth is provided by a gene that originated by overprinting. The lack of mutational studies on the overlapping region of genes A and C does not enable us to make inferences on the effects of gene overlap.

The proposed ancestry of genes B and D is consistent with their function during the phage life cycle. Both genes encode essential structural proteins that are required for the phage procapsid to be formed (Dokland et al., 1997).

Positive selection found in gene B is consistent with the detection of a big-benefit mutation that allows phage growth at high temperature (Bull et al., 2000).

Although the presence of overlapping genes in coliphages has long been identified, studies on their origins are rather fragmentary. The method presented here is based on a detailed analysis of the codon-usage pattern. It takes advantage of the use of PCA, a multivariate statistical technique capable of providing a low-dimensional representation for large amounts of data. As shown in Fig. 1, the first two axes of ordination are more than sufficient to detect the main patterns of codon usage. By this approach, the increase in the genome complexity during coliphage evolution can be appreciated (Fig. 2).

The correlation test (between an individual overlapping gene and the entire set of non-overlapping genes) should be considered as an auxiliary tool, to meet the objection that the use of synonyms in short genes can be affected by biased amino acid composition. More generally, the utility of our method lies in the fact that it overcomes the need for a phylogenetic analysis. Thus, it could be an adequate tool for investigating the origins of gene overlap, especially in those viruses with poor phylogenetic information.

Acknowledgements

The author is grateful to Franco Conterio for his support and encouragement. A special thank you goes to Adriano Cosenza for preparing the figures and to Maria Cristina Cignatta for language revision. This work was financed by the MURST (Ministero dell’Università e della Ricerca Scientifica e Tecnologica).

References


Evolution of overlapping genes


