Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 9489–9493, October 1992
Evolution

Origins of genes: "Big bang" or continuous creation?

(overlapping genes/new genes/thyroid hormone receptor α2/plant viruses/human immunodeficiency virus)

PAUL K. KEESE* AND ADRIAN GIBBS†

*Commonwealth Scientific and Industrial Organisation, Division of Plant Industry, and †Research School of Biological Sciences, Australian National University, Canberra, ACT 2601, Australia

Communicated by Frank Fenner, May 28, 1992 (received for review February 21, 1992)

ABSTRACT **Many protein families are common to all cellular organisms, indicating that many genes have ancient origins. Genetic variation is mostly attributed to processes such as mutation, duplication, and rearrangement of ancient modules. Thus it is widely assumed that much of present-day genetic diversity can be traced by common ancestry to a molecular "big bang." A rarely considered alternative is that proteins may arise continuously *de novo*. One mechanism of generating different coding sequences is by "overprinting," in which an existing nucleotide sequence is translated *de novo* in a different reading frame or from noncoding open reading frames. The clearest evidence for overprinting is provided when the original gene function is retained, as in overlapping genes. Analysis of their phylogenies indicates which are the original genes and which are their informationally novel partners. We report here the phylogenetic relationships of overlapping coding sequences from steroid-related receptor genes and from tymovirus, luteovirus, and lentivirus genomes. For each pair of overlapping coding sequences, one is confined to a single lineage, whereas the other is more widespread. This suggests that the phylogenetically restricted coding sequence arose only in the progenitor of that lineage by translating an out-of-frame sequence to yield the new polypeptide. The production of novel exons by alternative splicing in thyroid receptor and lentivirus genes suggests that introns can be a valuable evolutionary source for overprinting. New genes and their products may drive major evolutionary changes.**

Just as the present universe arrived with a big bang, so too the fossils of the Burgess Shales record the abrupt appearance of a bewildering array of metazoan animals 520 million years ago (1). No qualitatively new structural body plan appears to have arisen since the Cambrian.

The widespread occurrence of related biological macromolecules and biosynthetic pathways that are common to all cellular organisms (2) suggests that there may have been a genetic equivalent of the metazoan explosion. Even genes that are phylogenetically restricted to recent groups of organisms, such as the haptoglobins of mammals, appear to be derived from more ancient genes—namely, serine proteases (3).

The "big bang" paradigm of molecular evolution has been reinforced by x-ray crystallography studies that reveal common structures shared by proteins that no longer share discernible amino acid sequence similarity (4). In addition, Dorit *et al.* (5) have suggested that all proteins are derived from a limited set of 1–7000 exons. Thus most present-day molecular diversity is commonly ascribed to factors such as mutation, DNA duplications and rearrangements, exon shuffling, transposition, and alterations to regulatory pathways (see ref. 6 and references therein). The possibility that entire genes or coding domains are created, *de novo*, and that this is a significant evolutionary process seems to have been largely ignored.

New genes can be generated in two different ways. Polynucleotide molecules can be polymerized, *de novo* (7, 8), or they can be generated by the translation of previously unused reading frames of existing coding and noncoding genomic material. The possibility of generating new genes from preexisting nucleotide sequences was first mooted by Grassé (9), who called it "overprinting," and more recently by Ohno (10).

New genes or coding regions that arise by overprinting are most readily detected in overlapping genes where the nucleotides providing the new gene already encode a gene whose function has been maintained. The first such overlapping genes were identified in the genome of ϕX174 in which some parts of a single nucleotide sequence are translated from two or even three different reading frames, thus giving rise to unrelated polypeptides (11, 12). Similar overlapping genes whose coding regions are translated in different reading frames or from complementary strands have subsequently been described for many viral and cellular genes. In some cases a sufficient number of related genes have been reported, which allow construction of phylogenetic histories distinguishing the order of appearance within each set of overlapping coding regions. Here we describe several examples of such cases and identify some of the evolutionary implications for *de novo* origins of coding sequences.

MATERIALS AND METHODS

The sequences of the ligand-binding domains of 40 members of the steroid/thyroid receptor superfamily and of the methyltransferase-like domain of 14 viruses were aligned, and their FJD distances were calculated by the progressive alignment program (13). The relationships of the eight β-strands in the β-barrel domain of the virion proteins of 11 viruses were obtained from the program DISTANCES (Genetics Computer Group; version 6, with a match threshold of 1.0; ref. 14). The neighbor-joining method (15) was used to calculate dendrograms from distances.

RESULTS

Overlapping Cellular Genes. Overlapping cellular genes, expressing unrelated proteins, have been reported for bacterial (16–19), mitochondrial (20), and nuclear (21–25) genes. Perhaps the clearest evidence of novelty resulting from genetic overprinting is in the thyroid hormone receptor (TR) α gene, which belongs to a family of genes that encode nuclear receptors whose ligands include steroids, vitamin D, retinoic acid, and thyroid hormones (26). The two most conserved functional domains bind DNA and hormone. The gene for TRα has two alternatively spliced forms in human

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: TR, thyroid hormone receptor; RP, replicase protein; OP, overlapping protein; ORF, open reading frame; HIV, human immunodeficiency virus.

(23) and rat (24). The first eight exons are common to both forms, but exon 9 is unique to TRα1, whereas exon 10 is unique to TRα2 (Fig. 1). The NH_2-terminal 370 amino acids are the same in both forms and include the DNA-binding domain. They differ in sequence and length at their COOH termini, which functions as part of the ligand-binding domain. The distinctive COOH terminus of TRα1 is 40 amino acids long and is able to bind triiodothyronine, whereas that of TRα2 has an unrelated 120 amino acid COOH terminus. TRα2 does not bind triiodothyronine but acts as a dominant negative regulator of TRα1 and TRβ expression. The 5′ terminal 263 nucleotides of the 360 encoding the TRα2-specific sequence overlap the 3′ terminal exon of a related thyroid receptor gene (*ear-1*, ref. 23; REV-erbA, ref. 24), which is encoded in the complementary strand adjacent to the TRα gene (Fig. 1). The 3′ exon of *ear-1* encodes part of its ligand-binding domain and binds triiodothyronine, although weakly.

When the sequences of the ligand-binding domains of these receptors are aligned, it can be seen that (*i*) the COOH-terminal 40 amino acids specific to TRα1 have a clear sequence similarity to the receptors of other members of the family; (*ii*) the region of the ligand-binding domain of *ear-1* whose nucleotide sequence is overlapped by the TRα2-specific sequence is also unequivocally related to ligand-binding domains of others of the family; and (*iii*) the COOH-terminal amino acids of TRα2 encoded by the nucleotides overlapping *ear-1* show no significant sequence similarity to the ligand-binding domains of other members of the family.

These features indicate that TRα1 and *ear-1* are the original genes and that exon 10 of TRα2 arose more recently, *de novo*. This hypothesis is supported by phylogenetic analysis of the ligand-binding domains (Fig. 2). No open reading frame (ORF) related to exon 10 of TRα2 could be detected in the closely related mouse peroxisome proliferator-activated receptor (a sister group) or in the human and rat vitamin D_3 receptor genes (the nearest outgroup). In addition, no related ORF could be detected in all other more distantly related receptor genes. TRα2 exon 10 appears to be a phylogenetically restricted gene overlapping *ear-1*. Therefore, it is unlikely that TRα2 exon 10 has an ancient origin and has been lost subsequently in every lineage except that of *ear-1*. Exon 10 is only reported for human and rat and is not apparent or is not transcribed in chicken (30) or *Xenopus laevis* (31). Thus TRα2 exon 10 may have arisen after the divergence of mammals from birds.

Overlapping Viral Genes. Some genes are common to a wide range of viruses. Many viral genes, however, are restricted to particular families or subfamilies. These include a range of overlapping genes that occur in viruses from about 50% of all families, including DNA and RNA viruses of both prokaryotes and eukaryotes (11, 12, 32–35). This suggests that overprinting may be common in viruses. Here we describe three sets of overlapping genes that are translated in different reading frames and whose separate origins can be clearly identified.

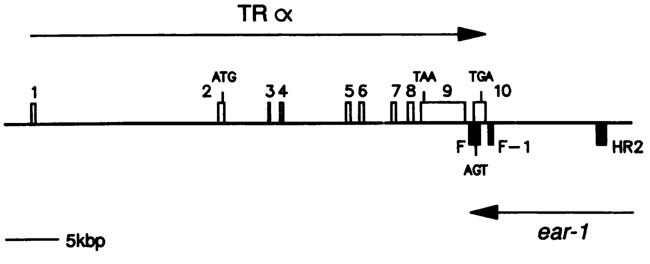

FIG. 1. Genomic map of the human TRα gene and part of the *ear-1* gene (27). Open boxes, TRα exons 1–10. Exons 9 and 10 are unique to TRα1 and TRα2, respectively. Black boxes, *ear-1* exons denoted F for final, F-1 for adjacent to final, and HR2 for homologue of rat exon 2. The arrows show the direction in which the genes are transcribed. All initiation and stop codons are shown except the initiation codon of *ear-1* (27).

The Overlapping Protein (OP) Genes of Tymoviruses. Over 90% of the tymoviral genome, including that of turnip yellow mosaic tymovirus, encodes a replicase protein (RP) of about 1800 amino acids, characterized by RNA replicase and helicase sequence motifs, and also a long OP that is necessary for viral spread (36) (Fig. 3*A*). The OP gene overlaps the 5′ terminal one-third of the RP gene, which encodes a protein with clear sequence similarity to the RNA methyltransferases of potexviruses, the closteroviruses, and Sindbis alphavirus (44).

The OP ORF is present in all tymoviruses that have been examined. It starts at the AUG that is closest to the 5′ terminus and is always separated from the AUG of the RP ORF by four nucleotides. However, no equivalent ORF overlaps the 5′ terminus of the RPs either of the closely related sister group of potexviruses or carlaviruses or of outgroups such as the alphaviruses and tobamoviruses (Fig. 4*A*). This suggests that the RP gene is the original gene and that the OP gene arose after the tymovirus and potexvirus RPs diverged from a shared ancestor. It is unlikely that this ORF was present earlier but was subsequently lost in all other related virus groups, while being maintained in every tymovirus.

Codon usage in the overlap regions supports the idea that the RP gene is the original gene. Tymovirus genomes have an unusually large proportion of cytosine residues, especially in the third codon positions of the RP and virion protein ORFs (37), and this bias is maintained in the portion of RP that overlaps with OP. The favoring of cytosine residues in the third codon position increases the chance that ORFs of significant length appear in alternative reading frames. Neither UAA, UAG, nor UGA stop codons contain cytosine. Large random ORFs may have been an important prerequisite for the appearance of the OP gene. Other large ORFs occur in tymoviral genomes, but none of these are found in all members or isolates of the group.

The M_r 17,000 Protein of Luteoviruses. An ORF encoding a protein of M_r of ≈17,000 is found embedded in the coat protein genes of all luteoviruses (Fig. 3*B*; ref. 39). Antibodies to the M_r 17,000 protein cross-react with the 5′ genome-linked protein that has a M_r of 17,000 but that is readily processed upon storage to a M_r 7000 form (ref. 45; R. R. Martin, personal communication). Sequence analyses of the luteovirus coat proteins show their close similarity to the eight-stranded β-barrel domains of most other virion proteins, which form small icosahedral particles (39, 46). Sequence comparisons show that the luteovirus coat proteins form a monophyletic group most closely related to the tomato bushy stunt virus/turnip crinkle virus/southern bean mosaic virus cluster (Fig. 4*B*). However, only the luteoviruses contain an overlapping ORF embedded within the coat protein gene. It is unlikely that the coat protein/17K (ORF encoding the M_r 17,000 protein) bifunctional gene arose first and that all other viral β-barrel coat proteins have subsequently lost the 17K homologue, both in the sister groups (tombusviruses, carmoviruses, and sobemoviruses) and the outgroups (picornaviruses, comoviruses, parvoviruses, and nodaviruses). As the 17K ORF homologue is present in the coat protein of all luteoviruses, it is more likely that the 17K gene arose *de novo* soon after the luteovirus coat protein gene diverged from those of other viruses.

The Lentivirus OPs. HIV-1 and -2 also have several overlapping ORFs (Fig. 3*C*), expressed mostly by alternative splicing (47). The 3′ terminal coding exons of the regulatory genes *tat* (48) and *rev* (49) overlap with one another and with the *env* gene; thus, all three reading frames in this region are

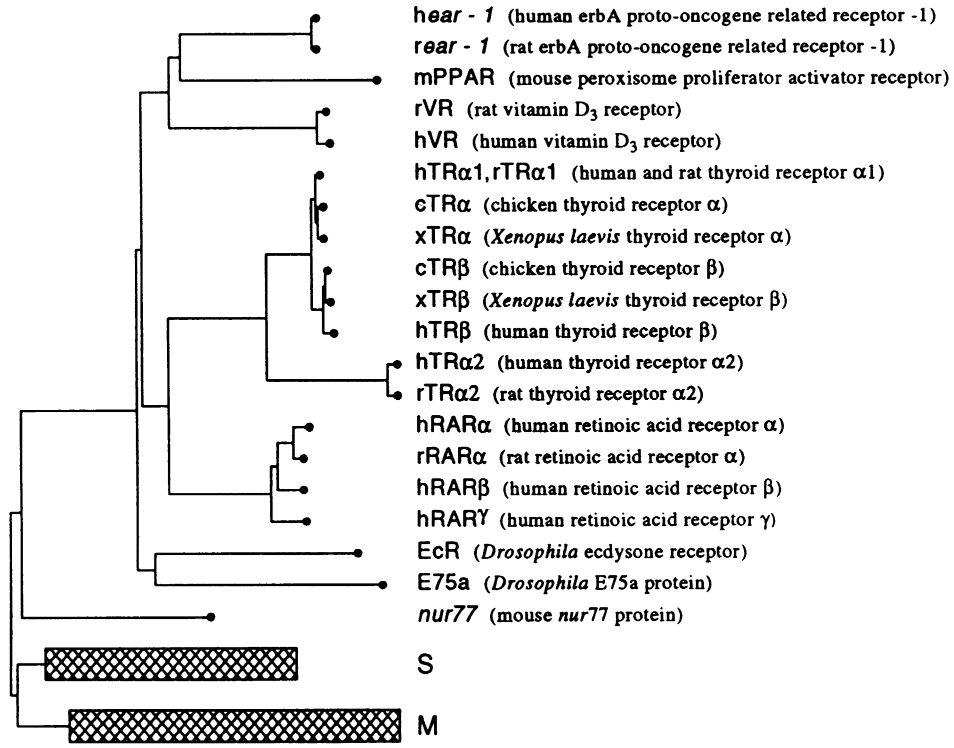
hear - 1 (human erbA proto-oncogene related receptor -1)
rear - 1 (rat erbA proto-oncogene related receptor -1)
mPPAR (mouse peroxisome proliferator activator receptor)
rVR (rat vitamin D3 receptor)
hVR (human vitamin D3 receptor)
hTRα1, rTRα1 (human and rat thyroid receptor α1)
cTRα (chicken thyroid receptor α)
xTRα (Xenopus laevis thyroid receptor α)
cTRβ (chicken thyroid receptor β)
xTRβ (Xenopus laevis thyroid receptor β)
hTRβ (human thyroid receptor β)
hTRα2 (human thyroid receptor α2)
rTRα2 (rat thyroid receptor α2)
hRARα (human retinoic acid receptor α)
rRARα (rat retinoic acid receptor α)
hRARβ (human retinoic acid receptor β)
hRARγ (human retinoic acid receptor γ)
EcR (Drosophila ecdysone receptor)
E75a (Drosophila E75a protein)
nur77 (mouse nur77 protein)
S
M

FIG. 2. Dendrogram of the ligand-binding domains (26–29) of steroid receptors. Branch S consists of several steroid receptors including glucocorticoid, androgen, progesterone, mineralocorticoid, estrogen, and estrogen-related type 1 and 2 receptors. Branch M consists of other receptors including chicken ovalbumin upstream promoter "orphan receptor," human retinoid receptor, and proteins from *Drosophila* seven-up, tailless, ultraspiracle, and *FTZ-F1* genes. A dendrogram with similar relationships has been reported by Laudet *et al.* (29).

expressed. Although the *env* genes of HIV-1 and -2 are related to the *env* genes of all other retroviruses (50), it is only the lentiviruses that have the *tat* and *rev* genes. Thus, it appears that the *env* gene is ancestral and the *tat* and *rev* genes arose later, *de novo*, in lentiviruses, including visna lentivirus, equine infectious anemia virus, and HIV-1 and -2 (see ref. 51). However, only HIV-1 and -2 (and the closely related simian immunodeficiency viruses) have the 3′ exon of *tat* that overlaps both the *rev* and *env* genes in the third reading frame (Fig. 3*C*).

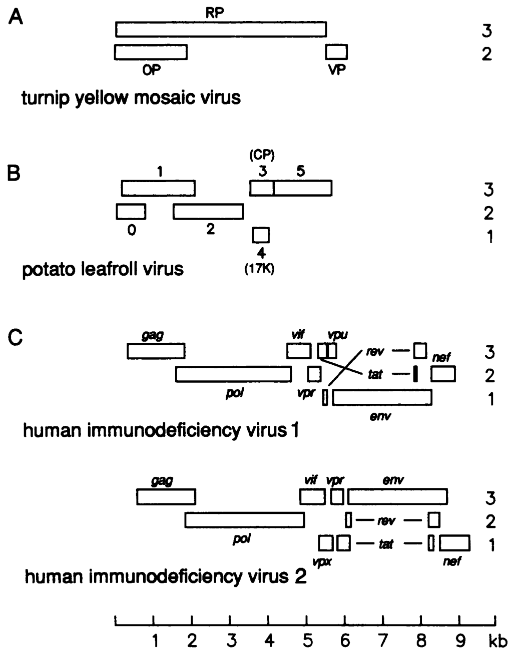

FIG. 3. (*A*) Viral genome map of turnip yellow mosaic tymovirus (37). VP, virion protein. (*B*) Viral genome map of potato leafroll luteovirus (38, 39). CP, coat protein; 17K, ORF encoding a M_r 17,000 polypeptide. (*C*) Viral genome map of the lentiviruses human immunodeficiency virus (HIV) type 1 and type 2 (40–43). tat, transactivator; rev, regulator of expression of virion proteins. kb, kilobase(s).

The HIV-1 genome also contains an additional gene, *vpu*, that overlaps the *env* gene but is not present in HIV-2 (Fig. 3*C*; ref. 42). Similarly, *vpx*, which overlaps the *vif* and *vpr* genes, is found only in the HIV-2 genome (52). As *vpu* and *vpx* are not found in HIV-2 and -1, respectively, or in other retroviruses, then they too probably arose *de novo* recently.

The *rev, tat, vpu,* and *vpx* genes of lentiviruses are all involved in higher order regulatory functions (see refs. 51 and 53). It appears that *rev* arose *de novo* in the lentiviruses after divergence from other retrovirus groups. The 3′ exon of *tat* then arose in HIV-1- and -2-related viruses after divergence from visna lentivirus and equine infectious anemia virus. Finally, *vpu* and *vpx* appeared after divergence of the type 1 and 2 forms of simian immunodeficiency virus and HIV. It is interesting that, when tested experimentally, *vpu* and *vpx* are found to be dispensable, whereas *rev* is not (54), because one would anticipate that new genes would be dispensable at first, but would become essential after selection and adaptation.

DISCUSSION

The phylogenetic histories of the TRα2 exon 10 and viral overlapping genes are the strongest evidence for differentiating new and ancient coding regions. Unequal crossing-over, slippage, gene conversion, RNA-mediated recombination, transposition, and mutation have contributed to genetic variation of ancient genetic segments: overprinting can generate genetic novelty. The view of genomes as dynamic systems is reinforced by the flexible and complex variations in expression patterns, such as ribosomal frameshifting, read-through of stop codons, ribosomal choice of AUG and non-AUG initiation codons, alternative splicing in number and order of exons, and RNA editing. These mechanisms, which are used for the expression of overlapping coding regions (e.g., ref. 55), can expose alternative ORFs to selection of novel functions.

If overlapping genes reflect the creation of new coding sequences, then several corollaries follow.

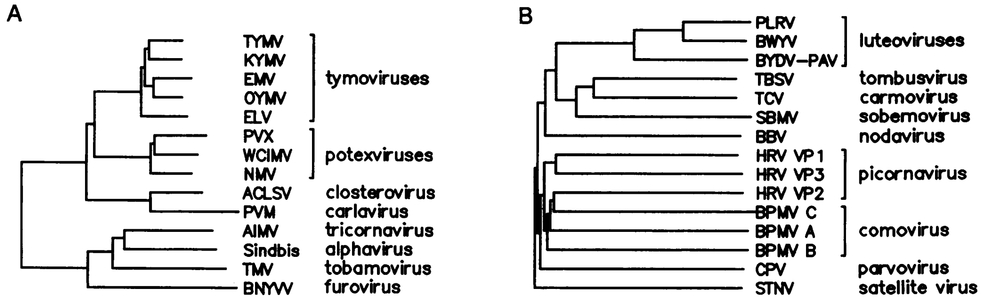

FIG. 4. (*A*) Dendrogram of the methyltransferase-like domain of the RPs of 14 Sindbis-like viruses. TYMV, turnip yellow mosaic virus; KYMV, kennedya yellow mosaic tymovirus; EMV, eggplant mosaic tymovirus; OYMV, ononis yellow mosaic tymovirus; ELV, erysimum latent tymovirus; PVX potato X potexvirus; WClMV, white clover mosaic potexvirus; NMV, narcissus mosaic potexvirus; ACLSV, apple chlorotic leaf spot closterovirus; PVM, potato M carlavirus; AlMV, alfalfa mosaic alfamovirus; TMV, tobacco mosaic tobamovirus; BNYVV, beet necrotic yellow vein furovirus. (*B*) Dendrogram of the virion proteins. PLRV, potato leafroll luteovirus; BWYV, beet western yellows luteovirus; BYDV-PAV, PAV strain of barley yellow dwarf luteovirus; TBSV, tomato bushy stunt tombusvirus; TCV, turnip crinkle carmovirus; SBMV, southern bean mosaic sobemovirus; BBV, black beetle nodavirus; HRV, human rhinovirus; BPMV, bean pod mottle comovirus; CPV, canine parvovirus; STNV tobacco necrosis satellite virus.

(*i*) Many genes or exons may have arisen by overprinting. The maintenance of two functional overlapping genes is likely to constrain the ability of both genes to become optimally adapted. Such constraints are avoided by duplication of overlapping genes and allowing each gene to adapt optimally in one of the resulting copies. In eukaryotes, where gene duplication is common, useful overlapping genes will probably rapidly become separated in this way. Overlapping genes probably persist in viruses more often due to constraints on genome size.

(*ii*) Genes that have arisen by overprinting may be identified by their biased composition. Out-of-frame expression of a gene with strongly biased composition will move a bias in the third codon position into the first or second codon positions of the novel ORF. Thus the gene will have an unusual codon usage and encode new proteins with physicochemically biased properties. For example, those encoded by genes with a preponderance of cytosine in the first position, like the OP gene of the tymoviruses, will tend to be basic, whereas those from genes with an excess of thymine in the second position, like the E gene of ϕX174, will be hydrophobic (11).

(*iii*) All nucleotide sequences have redundant ORFs, which are potentially useful coding regions. Alternative splicing may be a common way in which these ORFs are expressed in eukaryotes and their viruses to produce new protein domains. The possibility that splicing changes could generate novel genetic information in eukaryotes was first suggested by Gilbert (56) but has since been overlooked despite many examples of homologous proteins with alternative splicing patterns (57). The complex splicing patterns of TR and lentivirus genes indicate how small mutational changes can produce new ORFs to give proteins with large functional differences. Interestingly, the average exon size of 35–45 amino acids (58) is not much greater than the average length of ORFs found in random unbiased nucleotide sequences (64 codons/3 stop codons = 21.3). The concatenation of ORFs derived at random can thus generate novel coding sequences.

In organisms that lack significant RNA splicing, the chance occurrence of longer ORFs is enhanced by a range of nucleotide biases. The UA doublet occurs significantly less often in all reading frames of both coding and noncoding sequences in most genomes (59). UA comprises the first two residues of the stop codons UAA and UAG. Other biases may be more species specific such as the large cytosine content in the third position of tymoviral ORFs. In addition, the codons UUA, UCA, and CUA are rarely used in highly expressed genes of *Escherichia coli* (60). These codons are complementary to the stop codons UAA, UGA, and UAG. It has therefore been noted that long, apparently nonfunctional ORFs occur in phase in the complementary strand of *E. coli* genes (61).

(*iv*) New genes may be important in the establishment and distinctive biology of different virus groups. Many overlapping genes correlate precisely with previously defined taxons. For example, the OP gene is present in all tymoviruses but no other viruses. Similarly, the M_r 17,000 protein of luteoviruses strictly correlates with members of that virus group. Thus the appearance of these novel genes may have provided the saltatory step in generating a new virus group.

(*v*) Cells may contain a pool of "junk" protein domains resulting from the translation of transient novel ORFs or by mutations in redundant gene segments. However, novel coding regions may become established in a population if they confer some advantage to those individuals containing them.

(*vi*) Some structural similarities could reflect evolutionary convergence. During evolution, proteins seem to retain their three-dimensional structural similarity much longer than the similarity of their amino acid sequences (4). Thus structural similarity is thought to be evidence of a common origin, even for proteins that show no sequence similarity. However, identical pentapeptide sequences can participate in either α-helices or β-strands (62), so some structural features may be analogous to the "universal attractors" of chaos theory (63). This idea could be tested by determining whether unequivocally new proteins from overlapping genes contain any of the characteristic structural folds or motifs of ancient proteins.

Thus we conclude that, in general terms, it is likely that organisms have two classes of genes: (*i*) ancient "housekeeping" genes, most of which predate the divergence of prokaryotes and eukaryotes (e.g., those encoding rRNAs, tRNAs, glycolytic enzymes, ribosomal proteins, and nucleotide biosynthetic enzymes) and (*ii*) younger new genes that are phylogenetically restricted and encode proteins that have diverse functions specialized to the current life-style of the organism in which they are found.

We predict that many of the novel genes have arisen by overprinting. Their formation is probably episodic and may contribute to the type of evolutionary pattern called "punctuated equilibria" (64).

We thank Dr. Georg Weiller for doing sequence alignments and Drs. Peter Rathjen, Peter Lockart, and David Penny for helpful comments on a "primitive" version of the manuscript.

1. Morris, S. C. (1989) *Science* **246,** 339–346.
2. Britten, R. J. & Davidson, E. H. (1971) *Q. Rev. Biol.* **46,** 111–138.

3. Kurosky, A., Barnett, D. R., Lee, T.-H., Touchstone, B., Hay, R. E., Arnott, M. S., Bowman, B. H. & Fitch, W. M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3388–3392.
4. Matthews, B. W. & Rossmann, M. G. (1985) *Methods Enzymol.* **115**, 397–420.
5. Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990) *Science* **250**, 1377–1382.
6. Langridge, J. (1991) *Molecular Genetics and Comparative Evolution* (Research Studies Press, Taunton, Somerset, England), p. 216.
7. Kavaler, J., Davis, M. M. & Chien, Y. (1984) *Nature (London)* **310**, 421–423.
8. Biebricher, C. K., Eigen, M. & Luce, R. (1986) *Nature (London)* **321**, 89–91.
9. Grassé, P.-P. (1977) *Evolution of Living Organisms* (Academic, New York), p. 297.
10. Ohno, S. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2421–2425.
11. Barrell, B. G., Air, G. M. & Hutchinson, C. A., III (1976) *Nature (London)* **264**, 34–41.
12. Shaw, D. C., Walker, J. E., Northrop, F. D., Barrell, B. G., Godson, G. N. & Fiddes, J. C. (1978) *Nature (London)* **272**, 510–515.
13. Feng, D.-F. & Doolittle, R. F. (1987) *J. Mol. Evol.* **25**, 351–360.
14. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
15. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
16. Thomas, C. M. & Smith, C. A. (1986) *Nucleic Acids Res.* **14**, 4453–4469.
17. Rak, B., Lusky, M. & Hable, M. (1982) *Nature (London)* **297**, 124–128.
18. Thisted, T. & Gerdes, K. (1992) *J. Mol. Biol.* **223**, 41–54.
19. Barany, F., Slatko, B., Danzitz, M., Cowburn, D., Schildkraut, I. & Wilson, G. C. (1992) *Gene* **112**, 91–95.
20. Fearnley, I. M. & Walker, J. E. (1986) *EMBO J.* **5**, 2003–2008.
21. Jankowski, J. M., Krawetz, S. A., Walczyk, E. & Dixon, G. H. (1986) *J. Mol. Evol.* **24**, 61–71.
22. Adelman, J. P., Bond, C. T., Douglass, J. & Herbert, E. (1987) *Science* **235**, 1514–1517.
23. Miyajima, N., Horiuchi, R., Shibuya, Y., Fukushige, S., Matsubara, K., Toyoshima, K. & Yamamoto, T. (1989) *Cell* **57**, 31–39.
24. Lazar, M. A., Hodin, R. A., Darling, D. S. & Chin, W. W. (1989) *Mol. Cell. Biol.* **9**, 1128–1136.
25. Vellard, M., Soret, J., Sureau, A. & Perbal, B. (1991) *C. R. Acad. Sci. Paris* **313**, 591–597.
26. Evans, R. M. (1988) *Science* **240**, 889–895.
27. Issemann, I. & Green, S. (1990) *Nature (London)* **347**, 645–650.
28. Segraves, W. A. (1991) *Cell* **67**, 225–228.
29. Laudet, V., Hänni, C., Coll, J., Catzeflis, F. & Stéhelin, D. (1992) *EMBO J.* **11**, 1003–1013.
30. Forrest, D., Sjöberg, M. & Vennström, B. (1990) *EMBO J.* **9**, 1519–1528.
31. Yaoita, Y., Shi, Y. & Brown, D. D. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 7090–7094.
32. Beremand, M. N. & Blumenthal, T. (1979) *Cell* **18**, 257–266.
33. Kozak, M. (1986) *Cell* **47**, 481–483.
34. Samuel, C. E. (1989) *Prog. Nucleic Acids Res. Mol. Biol.* **37**, 127–153.
35. Fields, B. N. & Knipe, D. M., eds. (1990) *Fields Virology* (Raven, New York), p. 2336.
36. Bozarth, C. S., Weiland, J. J. & Dreher, T. W. (1992) *Virology* **187**, 124–130.
37. Keese, P., Mackenzie, A. & Gibbs, A. (1989) *Virology* **172**, 536–546.
38. Keese, P., Martin, R. R., Kawchuk, L. M., Waterhouse, P. M. & Gerlach, W. L. (1990) *J. Gen. Virol.* **71**, 719–724.
39. Martin, R. R., Keese, P. K., Young, M. J., Waterhouse, P. M. & Gerlach, W. L. (1990) *Annu. Rev. Phytopathol.* **28**, 341–363.
40. Ratner, L., Haseltine, W., Patarca, R., Livak, K. J., Starcich, B., Josephs, S. F., Doran, E. R., Rafalski, J. A., Whitehorn, E. A., Baumeister, K., Ivanoff, L., Petteway, S. R., Jr., Pearson, M. L., Lautenberger, J. A., Papas, T. S., Ghrayeb, J., Chang, N. T., Gallo, R. C. & Wong-Staal, F. (1985) *Nature (London)* **313**, 277–284.
41. Guyader, M., Emerman, M., Sonigo, P., Clavel, F., Montagnier, L. & Alizon, M. (1987) *Nature (London)* **326**, 662–669.
42. Cohen, E. A., Terwilliger, E. F., Sodroski, J. G. & Haseltine, W. A. (1988) *Nature (London)* **334**, 532–534.
43. Gallo, R., Wong-Staal, F., Montagnier, L., Haseltine, W. A. & Yoshida, M. (1988) *Nature (London)* **333**, 504.
44. Srifah, P. (1991) PhD. Thesis (The Australian National University, Canberra, Australia).
45. Mayo, M. A., Barker, H., Robinson, D. J., Tamada, T. & Harrison, B. D. (1982) *J. Gen. Virol.* **59**, 163–167.
46. Rossmann, M. G. & Johnson, J. E. (1989) *Annu. Rev. Biochem.* **58**, 533–573.
47. Schwartz, S., Felber, B. K., Benko, D. M., Fenyö, E.-M. & Pavlakis, G. N. (1990) *J. Virol.* **64**, 2519–2529.
48. Arya, S. K., Guo, C., Josephs, S. F. & Wong-Staal, F. (1985) *Science* **229**, 69–73.
49. Sodroski, J., Goh, W. C., Rosen, C., Dayton, A., Terwilliger, E. & Haseltine, W. (1986) *Nature (London)* **321**, 412–417.
50. McClure, M. A., Johnson, M. S., Feng, D.-F. & Doolittle, R. F. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2469–2473.
51. Cullen, B. R. (1991) *Annu. Rev. Microbiol.* **45**, 219–250.
52. Hu, W., Heyden, N. V. & Ratner, L. (1989) *Virology* **173**, 624–630.
53. Vaishnav, Y. N. & Wong-Staal, F. (1991) *Annu. Rev. Biochem.* **60**, 577–630.
54. Malim, M. H., Böhnlein, S., Hauber, J. & Cullen, B. R. (1989) *Cell* **58**, 205–214.
55. Curran, J., Boeck, R. & Kolakofsky, D. (1991) *EMBO J.* **10**, 3079–3085.
56. Gilbert, W. (1978) *Nature (London)* **271**, 501.
57. Breitbart, R. E., Andreadis, A. & Nadal-Ginard, B. (1987) *Annu. Rev. Biochem.* **56**, 467–495.
58. Traut, T. W. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2944–2948.
59. Nussinov, R. (1984) *Nucleic Acids Res.* **12**, 1749–1763.
60. Sharp, P. M. (1985) *Nucleic Acids Res.* **13**, 1389–1397.
61. Cascino, A., Cipollaro, M., Guerrini, A. M., Mastrocinque, G., Spena, A. & Scarlato, V. (1981) *Nucleic Acids Res.* **9**, 1499–1518.
62. Argos, P. (1987) *J. Mol. Biol.* **197**, 331–348.
63. Crutchfield, J. P., Farmer, J. D., Packard, N. H. & Shaw, R. S. (1986) *Sci. Am.* **255**, 38–49.
64. Eldredge, N. & Gould, S. J. (1972) in *Models in Paleobiology*, ed. Schopf, T. J. M. (Freeman, San Francisco), pp. 82–115.