Institut Pasteur
Microbes and Infection 16 (2014) 96–103
www.elsevier.com/locate/micinf

Review

Implications of HIV RNA structure for recombination, speciation, and the neutralism-selectionism controversy

Donald R. Forsdyke

Department of Biomedical and Molecular Sciences, Queen's University, Kingston, ON K7L3N6, Canada

Received 22 August 2013; accepted 24 October 2013
Available online 7 November 2013

Abstract

The conflict between the needs to encode both a protein (impaired by non-synonymous mutation), and nucleic acid structure (impaired by synonymous or non-synonymous mutation), can sometimes be resolved in favour of the nucleic acid because its structure is critical for a selectively advantageous genome-wide activity – recombination. However, above a sequence difference threshold, recombination is impaired. It may then be advantageous for new species to arise. Building on the work of Grantham and others critical of the neutralist viewpoint, heuristic support for this hypothesis emerged from studies of the base composition and structure of retroviral genomes. The extreme enrichment in the purine A of the RNA of human immunodeficiency virus (HIV-1), parallels the mild purine-loading of the RNAs of most organisms, for which there is an adaptive explanation – immune evasion. However, human T cell leukaemia virus (HTLV-1), with the potential to invade the same host cell, shows extreme enrichment in the pyrimidine C. Assuming the low GC% HIV and the high GC% HTLV-1 to share a common ancestor, it was postulated that differences in GC% had arisen to prevent homologous recombination between these emerging lentiviral species. Sympatrically isolated by this intracellular reproductive barrier, prototypic HIV-1 seized the AU-rich (low GC%) high ground (thus committing to purine A rather than purine G). Prototypic HTLV-1 forwent this advantage and evolved an independent evolutionary strategy – similar to that of the GC%-rich Epstein-Barr virus – profound latency maintained by transcription of one purine-rich mRNA. The evidence supporting these interpretations is reviewed.

Keywords: Base composition; Epstein-Barr virus; Immune evasion; Purine-loading; Retrovirus

1. Introduction

A recent report by Lawrie et al. [1] of genome-wide strong negative (purifying) selection at synonymous (non-amino acid-determining) sites in the fruit fly, marks the continuing, but slow, convergence of the ideas of population geneticists with those of biochemists and biophysicists. Prior to the emergence of nucleic acid sequencing technology in the late 1970s, the latter groups had, through studies of base composition, equilibrium gradient sedimentation and thermal denaturation profiles, demonstrated the segmentation of genomes into units of relatively uniform base composition (GC%), which might relate either to genes ("homostability regions" or "microisochores"), or to larger regions potentially encompassing many genes ("isochores" or "macroisochores") [2,3]. Departing from the view of many population geneticists that the evolutionary pressure on synonymous sites was selectively neutral [4,5], it was proposed that synonymous sites were functional and subjected to selection. For example, in 1976 Japanese biophysicists suggested [6]:

> "It is quite plausible … that the homostability region plays an important part somewhere in the biological process within which the DNA is closely related. If so, then the evolutionary selective force can be considered to have fixed such regions in DNA. From the size of the homostability region, recombination might be one possible process which is aided by it. In any case, the wobble bases [at synonymous sites] must give the necessary redundancy to make a

E-mail address: forsdyke@queensu.ca.

1286-4579/$ - see front matter © 2013 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.micinf.2013.10.017

homostability region without spoiling the biological meaning of the genetic code: the activity of proteins."

Consistent with this, analyses of the first nucleic acid sequences to become available led Richard Grantham to his "genome hypothesis" [7]. Here he proposed the profound functional significance of the genome *as a whole*, rather than of the genes it carried through the generations: "We conclude that mRNA sequences contain other information than that necessary for coding proteins. This other 'genome-type' information is mainly in the degenerate bases of the sequence. Consequently, it is largely independent of the amino acids coded." That the information obtained from protein-encoding parts of genes was indeed likely to be of 'genome type' became evident when the GC% of sequences flanking genes was found to correlate closely with that of synonymous codon positions [8,9]. In 1986 Grantham et al. [10] imagined a shortly impending convergence of neutralist and selectionist viewpoints:

"Although codon use is a characteristic of the genotype, most evolutionary analyses have been based on the phenotype. How much independence exists between the two levels of evolution has not been determined, although neutralists and selectionists are converging, which should help to find a solution. Possibly, future data on the relative rates of silent and non-silent mutations will help to clarify this situation."

However, Grantham, although suspecting it was nucleic acid *structure* that was conserved at synonymous sites, focussed only on mRNA structure. Indeed, decades later, while affirming that "the underlying biological function disrupted by these [synonymous] mutations is unknown, but it is not related to the forces generally believed to be … shaping the evolution of synonymous sites," Lawrie et al. [1] still held that "a strong possibility remains that the function underlying the strong constraint at synonymous sites is related to mRNA structure." Likewise, in 2013 Lind and Andersson [11] concluded that "the deleterious effects of synonymous mutations are not generally due to codon usage effects, but that mRNA secondary structure, is a major fitness constraint," and Park et al. [12] claimed "a major role of natural selection at the mRNA level in constraining protein evolution."

Of course, structure conservation at the mRNA level means that conservation is also at the DNA level of the corresponding genes. Which is primary and which is secondary? Since, the potential of duplex DNA to adopt stem-loop configurations is often better developed in non-genic regions and introns, than in exons [13–16], the pressure for conservation could have arisen primarily at the genome level (be the genome itself DNA or RNA). Indeed, in 1986 Bernardi and Bernardi [8] referred to selection acting on the "genome phenotype" rather than on the conventional phenotype. This would be consistent with a selective pressure relating to recombination [6]. Mutations likely to affect recombination are of two general forms, those affecting the proteins that mediate recombination (conventional phenotype), and those affecting their targets – DNA itself (genome phenotype). The former mutations would be localized to genes (mainly non-synonymous mutations), the latter would be genome-wide (synonymous and non-synonymous when within protein-encoding regions), and could involve DNA structure.

The forces manifest when such mutations occur, should also operate in organisms with RNA genomes. Because of high mutation rates and high clinical interest, studies of retroviral evolution now provide well-documented exemplars of the general forces operating on genome sequences [17]. These forces can either promote or decrease recombination. In the latter case, new species can arise. While mechanisms of speciation are controversial, there is general agreement that "mutation is crucial in speciation because reproductive barriers cannot be generated without mutations" [18]. But, for RNA viruses in general, Simon-Loriere and Holmes [19] have argued that "there is little evidence that recombination is favoured by natural selection to create advantageous genotypes or purge deleterious mutations, as predicted if recombination functions as a form of sexual reproduction." However, as reviewed here, at least in the case of HIV-1, studies of RNA structure and base composition suggest otherwise.

2. Recombination, rejuvenation and conservation of RNA structure

At its most fundamental level, sex is recombination [20,21]. The conjugation required for the meeting of paternal and maternal human gametes is an elaborate preliminary to the final meiotic meeting of parental genomes in the gonads of their offspring, where recombination occurs. The early microscopists who witnessed this final meeting described it as a "conjugation of the chromosomes" that was necessary for a rejuvenating "interchange of substances" [22]:

"The conjugation of the chromosomes in the synapsis stage may be considered the final step in the process of conjugation of the [parental] germ cells. It is a process that effects the rejuvenation of the chromosomes; such rejuvenation could not be produced unless chromosomes of different parentage joined together, and there would be no apparent reason for chromosomes of like parentage to unite."

In contrast to humans, retroviral sex is quite elementary [23]. Yet, like humans, retroviruses are diploid. This diploidy in HIV-1 can be heterozygous due to the viral strategy of mutation to near oblivion, so countering host defences. The degree of heterozygosity in two HIV genomes that are copackaged within an infectious retrovirus particle, if below the sequence difference threshold above which recombination is inhibited (see later), will allow viral rescue by recombination. Thus, within the next host cell, the two partially crippled genomes can repeatedly recombine to generate a rejuvenated form that will successfully colonize the vulnerable population of host cells (usually T4 lymphocytes) [24]. This implies that the virus will have accepted mutations supporting the ability to