Repeated Intragenome “Parasites” as a Factor in Molecular Coevolution

S. N. Rodin, Y. G. Matushkin, and J. S. Krushkal[1]

Introduction

Any genome, except for its presently neutral DNA (i.e. without coding sequence), comprises a perfect ensemble of functional genetic units, the range of these units having its roots in individual exons and being crowned by the most complicated supergene complexes. The whole ensemble is undoubtably a product of mutually adaptive molecular coevolution. Any ecosystem, in turn, is the result of concerted molecular evolution of the species making up the system. At present, when the sequencing of entire genomes is running at a phenomena rate, to construct a theory of molecular coevolution would be of utmost importance both for theoretical molecular biology and genetics and for evolutionary theory itself.

All specific forms of adaptive molecular coevolution may be subdivided into intra- and intergenome and into directed and nondirected processes [1].

It is typical of nondirected molecular coevolution that, whatever its mode, mutations occur and are fixed at a rate which, despite their apparent adaptive value, remains on average constant. This fact, however, is at variance with one of the keystone postulates of Kimura's neutralistic theory. Up until now, only nondirected processes of molecular coevolution have been proposed and studied in details: intergenomically, concerted fixations of mutant reception and absorption genes in bacteria and phages, respectively [2], different variants of coevolving antigens and antibodies [3], original interactions between natural selection and molecular drive in the coevolution of multiple promoter and enhancer regions in rDNA loci, on the one hand, and the RNA *PolI* gene, on the other [4], specific pairs of base substitutions compensating for each other to maintain the rRNA secondary structure [5], etc. All these cases of molecular co-evolution, both intra- and intergenome (in different parasite–host pairs), are in fact variations on a theme, i.e. coevolution. The question of whether coevolution could provide development of multigene systems ab simplecioribus ad complexiora is of especially profound interest.

Regarding the genes of the immune system, we have suggested that HIV-like viruses could be involved in coevolution of this sort [6].

Coevolutionarily Motivated Complication of Immune Multigene Families

Let us consider a hypothetical ancestral organism with a primitive, poorly differentiated immune system. Suppose the corresponding ancestral immune cells (prelymphocytes) change their state from L to T in the course of ontogenesis, where L and T are the immature and mature prelymphocytes, respectively. Suppose also that viruses (V) can only strike the L-cells, i.e. immature prelymphocytes. We then admit that the molecular-genetic system of immunity is simple enough for

[1] Institute of Cytology and Genetics, USSR Siberian Academy of Sciences, Novosibirsk, USSR.

Haematology and Blood Transfusion Vol. 35
Modern Trends in Human Leukemia IX
R. Neth et al. (Eds.)

the virus to make use, by adsorption, of the very receptor of the L-cell that T-cells, in turn, use to identify and inactivate the virus. The adsorption of V on L leads to the formation of infected cells (denoted Z), from which, via lysis, the daughter viral particles come. The T/V binding, by contrast, leads to the elimination of the viruses. Accordingly, we can derive the following system of differential equations describing the dynamics:

$$\begin{aligned}
\dot{L} &= [F(L) - \alpha L] - G(L,V)\\
\dot{V} &= [\omega\beta Z - G(L,V)] - Q(T,V)\\
\dot{Z} &= G(L,V) - \beta Z \qquad (1)\\
\dot{T} &= \alpha L - kT
\end{aligned}$$

where α, ω, β and k are constants of the respective process rates.

The state of equilibrium $(\dot{L},0,0,\dot{T})$ in Eq. 1, where $\dot{T} = (a/k)\dot{L}$ and $\dot{L}$ is the root of equation $F(L) = aL$, is assumed to be health. This state is locally stable (which implies that the prelymphoid tissue is resistant to minor infections) if $Q'_V(\dot{T},0) > (W-1)G'_V(\dot{L},0)$ and unstable if $Q'_V(\dot{T},0) < (W-1)G'_V(\dot{L},0)$. In our model, an increase in the stability of the "healthy" state can be obtained by increasing the value of the term $Q'_V(\dot{T},0)$ and/or decreasing $G'_V(\dot{L},0)$. However, since $\dot{T} = (a/k)\dot{L}$, then a drop in $\dot{L}$ causes a drop in $\dot{T}$. To increase $\dot{T}$, it is necessary to increase $\dot{L}$; note that $\dot{T} < (a/k)\dot{L}$.

Therefore, there are two ways of increasing the resistance of the prelymphoid system to minor infections (in terms of a simplified model): first, by increasing the number of clones of those prelymphocytes that are specific to various antigen determinants; and secondly, by changing in the course of the prelymphocytes' maturation the avidity of the antigen specific receptor. Both ways are found in the immune systems of contemporary vertebrates.

The second way actually implies that no entirely identical receptor molecules can participate in either the absorption of viruses upon the target immunocytes or in the recognition and destroying of viruses (or their antigens), since to acquire homologous but not identical receptors, progressive divergence of the molecular genetic system of immunity is required.

However, what factor(s) could direct the evolutionary complication of all the other multigene families (MFs)? What if the role of intragenome "parasites", such as human *Alu* repeated sequences, retroviruses, and mobile genes of *Drosophila*, in the evolution of MFs is similar to that of HIV-like pathogens in the evolution of immune supergenes?

Intragenome Parasites and Genome: a Coevolutionary Aspect

We have studied [7] the processes of concerted variability which actually result from cooperation of such entities as, on the one hand, various mobile elements (a kind of "intragenome parasite", GP) and, on the other, the genome itself ("host").

Several systems of differential equations similar to Eq. 1 have been built in order to analyse the following situations:

1) the GP is insertable in the vacant sites only, its free state (not in the "host" but still in the cell) not being durable;
2) the GP is insertable in the vacant sites only, its free state being durable;
3) the GP is insertable in both vacant and occupied sites („molecular memory"), its free state not being durable (mammalian *Alu*-like repeats taken as a prototype);
4) the GP is insertable in both vacant and occupied sites and is able to exist "on its own" (retroviruses taken as a prototype).

We then admitted that the genome is tolerant to the "selfish" proliferation of GP until the share of the occupied sites exceeds the limit $1/K$. Our analysis revealed that the coevolutionary complication of GP – from the simplest, which is only able to insert in vacant sites, through the ongoing acquirement of terminal repeats ("molecular memory"), to perfectly integrated complexes with an extragenomial life style – is accompanied by change in the selective coevolutionary restrictions on genome size: upper limit–no limit–lower limit. Thus, mobile elements may be regarded as an inner factor inducing progressive, coevolutionarily motivated complication of genomes, including multiplication of coding regions.

Our models are based on the assumption that there is always a superior selective force (from the "host" side) that restricts the number of GPs and influences the pattern of GP distribution in the host genome. However, the following question arises here: Are there any inferior restrictions directly related to the GP structure as such? We go on to show below that *Alu*-like repeated sequences, even with extremely simple structures, could have such restrictions.

CpG-Rich Promoters as an Inner Constraint on Amplification of *Alu*-Like Sequences

With the aid of the package of applied programs VOSTORG [8], designed in our laboratory, 83 *Alu* repeats (60 human included) from seven species of primates and 13 *Alu*-like *B*1 repeats from three rodent species were subjected to phylogenetic analysis, in particular, for mutations fixed in RNA polymerase III promoters (Fig. 1).

Using the method of diagnostic positions [9] enabled us to divide all 60 human *Alu* sequences into three different classes (Fig. 2) corresponding to *J*, *Sa* and *Sb* (identification of the *Sc* class was certain) according to Britten et al. [9]. The topologies of the phylogenetic tree constructed on the complete sample of *Alu* sequences and of the tree derived from the comparison of the consensus for all classes revealed a good agreement with the order of appearance of these classes in the course of evolution (Fig. 3): progenitor (*7SL* RNA gene) → *J* → *Sa* → *Sc*(?) → *Sb*.

As is known, the CpG positions evolve on average 10.5 times as much as other positions of *Alu* repeats, which is due to methylation of the cytosines. In particular, the A (enhancing) and B (initiating) boxes of promoters contribute considerably to the concentration of CpGs (Fig. 1). We tried to build a dichotomic dendrogram from CpG positions of the promoter alone but failed. This could be an argument in favour of the "burst"-like formation of the *Alu* classes.

The most intriguing feature of the *Alu* evolutionary tree (Fig. 3c) is the almost absolute lack of mutations in CpG dinucleotides of the promoter region at the

```
          134  139        150
Right     AATTAGCCgGGcgTGGTGGcgcgCCTGTAATCCCAGCTACTcgGGAGGCTGAGGCAGGA
Left      ----GGCcgGGcgcgGTGGCTCAcgCCTGTAATCCCAGCACTTTGGGAGGCcgAGGcgGGc
              1  4          15

                          210                                          251
Right     GAATcgCTTGAACCcgGGAGGcgGAGGTTGCAGTGAGCcgAGATcgcgCCACTGCACTCCA
Left      gGATCACCTGAGGTCAGGAGTTcgAGA-------------------------------CCA
                          75                                          86

                                             284
Right     GCCTGGGcgACAGAGcgAGACTCcgTCTCAAAAAAAA
Left      GCCTGGCCAACATGGTGAAACCCcgTCTCTACTAAAA
                                             120
```

Fig. 1. Consensus of human *Alu* repeats [9] with the left and right halves of the sequences aligned. A and B boxes of the promoter region are *underlined*. The CpG dinucleotides are in *lower case letters*. The right promoter is likely to be inactive due to the relatively long inserted sequence

```
 62 C ...........................TA.................T.TT.TGT.AGTTT
 63 A ......C...T...................................G.GG.GG-GGGGGG
188 C ...............G...................A..........T.TTTTT..T.T.T
189 A .................................C.......T..GGG.GGGGG.G..GG-
194 A ...........................T...............GGG..GGG.GGG.GGGG
101 G ..........A...AAA...........A..............A.AAAAAA..AA.A.AA
106 A ...........................................G.GG.G.....GGGGG.
 94 C .........T.....T.....G.........T...........GGAG.GGGGGGGGGGGG
 70 G .........T.......A........C................C.TC.CCCCCCC.CCC.
 71 T .........-.........C.......................CGCC.CCCCCCC.CCCC
204 A .....-........................G............GGGGTGGGGGGGGG.GG
220 T ...-............C...C..................G....CCC.CCCCCCCC.CCG
233 A .......................G.....C............GTTTT.TTTTTTTT.TTC
275 T ..................-.................C.C...CCCCC..C.CCCC-CCCC
208 G ....-....A......A.A..A....A.T-.AAAA.A.T.A.AAAA-A.A..AA...AA
 57 C ..T.T.AA.T...T..TG.TT.TT.TATT.AT...A.TTTAAAA.A.AATTAAAAAA
163 A GG.G.T..G...G.G...C..G..G..G.G...G..G..G.GCGGGTGGCGGGGG.GG
153 C GGGGG...GGG.GGGTT.GTG.GG..G-.GT....GT.GTG..T..GT....TTTT..T
 65 C ------..-TT.TT-.-T-T..A.T-T-T-..GT..-.-...TTTT-TTTTT-TTTTTT
 66 T ------..-.....-.-.-.......-.-.......-.-.......-.....-......
 78 T AAAA.A..A.....A.A.A........A.A......A..A............C......
 88 G TTTT.TA.T.....T.T.T........T.T.C....T.............C.....A.
 95 C TTTT....T.....T.T.........T.T.................T..T.......
100 T CCCC....C....AC.C........C.C................A..A..C.......A
197 C GGGGT.G.G....AG......T...TTG.G..GT.T..TTT....T.....G...T...
200 T GAGA.-..G.....A............G............C..C......CC.......
219 G CCCC....C.....C-...........C.C..................A....A.....
      SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSJJJSJJJJJJJJJ
      bbbbacaabaaaaabacacaaaaaaaababaaaaaacaaaaa        a
```

Fig. 2. Variability in the diagnostic positions of 60 human *Alu* repeats. On the *left*, the consensus nucleotides are shown. All the *Alu* repeats rearranged in accordance with a divergence from the consensus. *Letters* at the *bottom* indicate the class to which each *Alu* repeat

upper branches of the tree (when the *Alu* subfamilies were being newly formed), in spite of their extreme mutability. Non-CpG positions show the same regularity (Fig. 3e). It should be noted that mutations in the CpG sites of the "quasi-neutral" part of the *Alu* repeats appear not to be in deficit at that period (Fig. 3d). Thus, a lion's share of mutations in the promoter CpG sites are concentrated in the lower branches of the divergency tree just after the class formation process is over (not shown).

This means that the promoter region of *Alu* repeat progenitors were under very strong negative natural selection pressure until the amplification process started. Moreover, the topology of the dichotomic branching within each class appears to be unstable. Thus, during evolution, first some changes in diagnostic positions (CpG sites not belonging to them) had to be accumulated; secondly, a current class of *Alu* sequences branched off the main stem of the tree; and, finally, mutations at CpG positions predominantly within the promoter occurred most rapidly. The superfamily of *B*1 repeats of rodents, closely related to *Alu*, shows similar regularities.

The results obtained allow us to propose a model where promoter sites play a role of profound importance both in intragenome amplification of the progenitor *Alu* sequence and in the divergence of individual members of the corresponding subfamilies. The model is supposed to explain the limited sizes of a subfamily with the subsequent acquisition of mutational defects in CpG positions of promoters and hence the inevitable slowdown of amplification. As a result, only those of *7SL* RNA-like sequences which have retained the promoters could

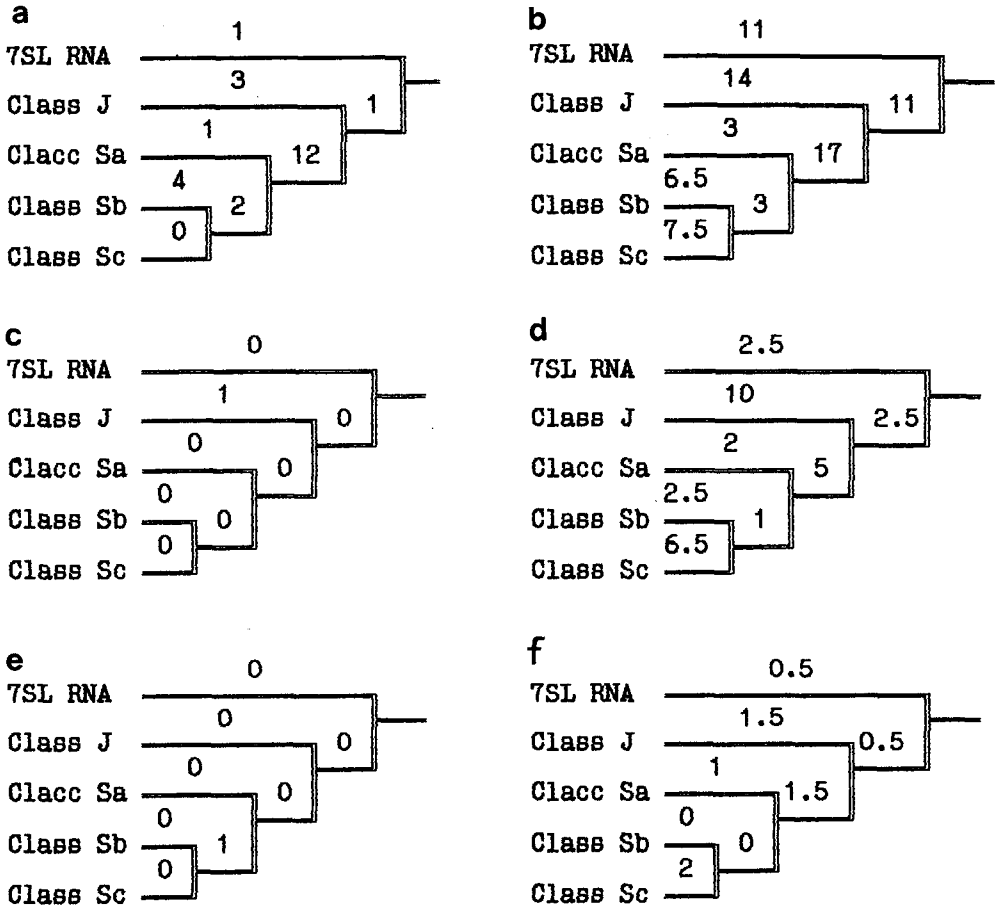

Fig. 3a–f. Phylogeny of the consensus sequences reconstructed for the main *Alu* classes with a human *7SL* RNA sequence as a repeat. Numbers of mutations fixed in various types of positions are shown: **a** in the 23 diagnostic non-CpG positions; **b** in all positions without central and terminal oligo-A parts; **c** in 8 CpG positions in A and B boxes of the left (active) promoter for the host RNA polymerase II; **d** in 38 non-promoter CpG positions; **e** in 16 non-CpG positions in the left promoter; **f** in 6 CpG positions from sites in the right (inactive) domain homologous to A and B boxes

become the progenitor for the following subfamily of *Alu* repeats to amplify and evolve in an active mode.

Each *Alu* repeat is well known to consist of two homologous halves (Fig. 1). Usually, only the leftmost domain is active for amplification by reverse transcription [10]. Figure 3f shows that, in contrast to the single CpG mutation in the leftmost promoter, the rightmost one accepted seven such mutations in CpG dinucleotides at the top part of the tree just when the *Alu* subfamilies were in the making. This is an additional, rather convincing, argument in favour of the importance of the promoter CpG sites, in particular those located in A and B boxes.

Thus, the "selfish" intragenome propagation of any progenitor "pregnant" with a recurrent *Alu* subfamily is destined to slow down and, eventually, to come to a standstill because as any individual *Alu* promoter rapidly accumulates more and more defects, predominantly due to the increased mutability of CpG sites, the host reverse transcriptase becomes less able to recognize the promoter.

This is not so with HIV-like retroviruses. They show unusually high variability, generated by viral reverse transcriptase, the most error-prone of the various RNA and DNA polymerases [11]. In contrast to the short *Alu* repeating unit, the HIV reverse transcriptase is encoded by its own *Pol* gene. It produces extremely frequent mutations in all regions of the viral genome, including in its own gene. Therefore, there is a good chance for promoter sites and reverse transcriptase to be involved in prolonged steady coevolution, based on the selection of pairs of substitutions compensating for each other. It is an original case of a strikingly rapid intragenome coevolution which should be adaptive but is apparently not directed.

Summary and Conclusions

"Parasitic" DNA may be regarded as a rather active partner in different coevolutionary processes. The basic stages of the processes are likely in most cases to be as follows: parasitism → tolerance → → symbiosis. There are interior and exterior coevolutionary factors complicating molecular-genetic systems within a supersystem "mobile elements-genome". For example, the data presented above indicate clearly that the relatively high concentration of CpG sites in the *Alu* promoter looks prudent as regards the needs of the "parasite" as well as those of the host genome. We consider "prudence" of this kind to be most likely a product of large-scale molecular coevolution.

As to HIV-like retroviruses, they could be simultaneously involved in three different regimes of molecular coevolution:

1) at a level of the parasitic genome as such;
2) as a typical intragenome parasite inserted in the host genome inducing complication in multigenic system (like *Alu*);
3) as a typical intracellular parasite in an "active", infectious state stimulating complication in the immune multigene families.

Evidently, it is only the steadiness of the first coevolutionary process (with "no wheels, no sails") provides for a possible role of HIV-like parasites as a selective factor provoking coevolutionary complication of host genomes.

References

1. Rodin SN, Rzhetsky AY, Matushkin YG (1987) Coevolutionary approach to the motivation of molecular-genetic organization of immune system. In: Mlikovsky J, Novak V (eds) Towards a new synthesis in evolutionary biology. CSAV, Prague, pp 130–132
2. Rodin SN, Ratner VA (1983) Some theoretical aspects of protein coevolution in the ecosystem "phage-bacteria". II. The deterministic model of microevolution. J Theor Biol 100:197–210
3. Rodin SN, Rzhetsky AY (1989) Coevolutionary approach to the problem of molecular-genetic bases of antibody diversity (in Russian). Achievements Contemp Biol 107:357–374
4. Dover GA, Flavell RB (1984) Molecular coevolution: DNA divergence and the maintenance of function. Cell 38:622–623
5. Hancock JM, Tautz D, Dover GA (1988) Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. Mol Biol Evol 5:393–414
6. Rodin SN, Matushkin YG (1987) Intracellular infections as one of the factors directing progressive divergency of immune multigene system in evolution (in Russian). J Gen Biol 48:845–856
7. Rodin SN (1991) Idea of coevolution. Chapman and Hall, London (in press)
8. Zharkikh AA, Rzhetsky AY, Morozov PS, Sitnikova TL, Krushkal JS, Matushkin YG (1990) VOSTORG: package of a microcomputer programs of phylogenetic analysis. Gene (in press)
9. Britten RJ, Baron WF, Stout DB, Davidson EH (1988) Sources and evolution of human *Alu* repeated sequences. Proc Natl Acad Sci USA 85:4770–4774
10. Perez-Stable C, Shen C-K (1986) Competitive and cooperative functioning of the anterior and posterior promoter elements of an *Alu* family repeat. Mol Cell Biol 6:2041–2052
11. Varmus H (1988) Retroviruses. Science 240:1427–1435