Horizontal Gene Transfer Contributes to the Wide Distribution and Evolution of Type II Restriction-Modification Systems

Albert Jeltsch, Alfred Pingoud

Institut für Biochemie, FB 15, Justus-Liebig Universität Giessen, Heinrich-Buff-Ring 58, 35392 Giessen, Germany

Received: 7 June 1995 / Accepted: 14 September 1995

Abstract. Restriction modification (RM) systems serve to protect bacteria against bacteriophages. They comprise a restriction endonuclease activity that specifically cleaves DNA and a corresponding methyltransferase activity that specifically methylates the DNA, thereby protecting it from cleavage. Such systems are very common in bacteria. To find out whether the widespread distribution of RM systems is due to horizontal gene transfer, we have compared the codon usages of 29 type II RM systems with the average codon usage of their respective bacterial hosts. Pronounced deviations in codon usage were found in six cases: EcoRI, EcoRV, KpnI, SphI, Smal, and TthHB81. They are interpreted as evidence for horizontal gene transfer in these cases. As the methodology is expected to detect only one-fourth to one-third of all horizontal gene transfer events, this result implies that horizontal gene transfer had a considerable influence on the distribution and evolution of RM systems. In all of these six cases the codon usage deviations of the restriction enzyme genes are much more pronounced than those of the methyltransferase genes. This result suggests that in these cases horizontal gene transfer had occurred sequentially with the gene for the methyltransferase being first acquired by the cell. This can be explained by the fact that an active restriction endonuclease is highly toxic in cells whose DNA is not protected from cleavage by a corresponding methyltransferase.

Key words: Restriction enzyme — Modification enzyme — Restriction endonuclease — DNA methyltransferase — Codon usage

Introduction

Type II restriction modification (RM) systems comprise two enzymes, a restriction endonuclease that specifically cleaves DNA within sequences 4–8 bp in length and a methyltransferase that specifically methylates the DNA within the same sequences at A or C residues and thereby protects these sequences from cleavage (reviews: Bickle and Knüfer 1993; Heitman 1993; Hornby 1993; Noyer-Weidner and Trautner 1993; Pingoud et al. 1993; Roberts and Halford 1993). Usually, cellular DNA is methylated and, hence, protected, whereas incoming DNA, for example, invading phage DNA, is not methylated and can be cleaved and inactivated. To date, more than 2,000 restriction endonucleases with approximately 200 different specificities have been identified (Roberts and Macelis 1994), although not all of them have been shown to be part of an RM system and only a minority of them has been sequenced. RM systems are widely distributed among bacteria (Wilson and Murray 1991). Their primary function seems to be the defense against foreign DNA in general and bacteriophages in particular. They seem to be very effective in this function, as can be deduced from the fact that based on crude estimates most bacteria harbor one or more RM system (Wilson and Murray 1991) and that for species with many RM systems, like Neisseria gonorrhoea, with 13 identified RM systems (Piekarsowicz and Stein 1995), no bacterio-
phages are known. To investigate the molecular evolution of type II RM systems, we recently have derived phylogenetic trees for restriction endonucleases and C(C5)- and A-methyltransferases. Although the amino acid similarities found are spurious, especially for the endonucleases, a comparison of the similarities within the amino acid sequences and the recognition sequences of the enzymes demonstrated that these trees are meaningful (Jeltsch et al. 1995). Surprisingly, none of these molecular trees resembles the phylogeny of bacteria (Olsen et al. 1994). For example, closely related enzymes, like the homologous isoschizomeric restriction enzymes EcoRI and RsrI (Stephenson et al. 1989), are found in bacterial species that are not closely related, namely, E. coli and Rhodobacter sphaeroides. Hence, the question arises of whether RM systems could have spread over the bacterial kingdom by horizontal gene transfer. To find out whether such a transfer has contributed to the distribution and evolution of type II RM systems, we have compared the codon usage of RM genes with the average codon usage of their respective bacterial hosts. If the codon usage of the genes for a restriction and/or modification enzyme deviates significantly from the codon usage characteristic for the host, this can be taken as evidence for a horizontal gene transfer. Taking into account that an active restriction endonuclease is toxic for cells not containing a protecting methyltransferase, one could envisage three possible ways horizontal gene transfer of RM systems might occur: (1) The genes for the endonuclease and methyltransferase are transferred en bloc, e.g., by heterologous transformation, (2) the genes are transferred one after the other, with the methyltransferase gene being acquired first, or (3) an inactive copy of the endonuclease gene is transferred and becomes reactivated after an appropriate methyltransferase gene is picked up. By comparing the codon usage deviations of the restriction endonuclease and the corresponding methyltransferase genes, conclusions can be drawn as to the temporal order of events—in particular, as to whether the two genes were transferred together or one after the other.

**Methods**

Codon usage frequencies and reference proteins for 41 different genera and species (Wada et al. 1991) were obtained from the EMBL database (Netser@embl-heidelberg.de). These data are compiled in Table 1. Sequences of genes for RM systems were obtained from the Genbank or EMBL data bank. All type II RM systems for which reference data are available have been included in the analysis, i.e., AccI, BamHI, BanI, BsuBI, BsuRI, CeqI, Eco47I, Eco57I, EcoRI, EcoRII, EcoRV, HincII, HpaI, HpaII, KpnI, NcoI, NsiI, PaeR7, PvuII, RsrI, SphI, Sau3AI, ScaI, SmaI, SstI and ThrHB81. Codon frequencies \(f\) were normalized for all codons encoding the same amino acid, e.g., \(f_{\text{ATT}}\) is normalized with respect to codons for isoleucine (ATT, ATC, and ATA): \(f_{\text{ATT}} = n_{\text{ATT}}(n_{\text{ATT}} + n_{\text{ATC}} + n_{\text{ATA}})\), where \(n_{\text{xyz}}\) represents the number of XYZ codons found in the gene.

The three stop codons are treated equally. Deviations in codon usage frequencies \(D\) are defined as the sum of squares of the 64 individual differences between the codon usage of codon XYZ in the protein under investigation and the average codon usage of all reference proteins: \(D = \sum (f_{\text{found}} - f_{\text{avg}})^2\). Codon usage deviations from the average codon usage calculated by this procedure are not influenced by the amino acid composition of the protein.

To evaluate the deviations found for the RM genes we calculated the codon usage deviations for all reference genes of the same species in the data base to determine the mean codon usage deviation \(\mu\) and its standard deviation \(\sigma\) (Table 1). Using these values a significance score \(Z\) is calculated: \(Z = (D - \mu)/\sigma\).

In the reference data base used, depending on the number of reference genes available, codon usages were calculated for genera or for single species. The computed mean codon usage deviations \(\mu\) and standard deviations \(\sigma\) are similar within a small range for all reference sets, regardless of whether single species or genera are considered (Table 1). In particular, standard deviations were not smaller in those cases where genes of single species are investigated as compared to those where genes of a whole genus are compared (e.g., \(\sigma[E. coli] = 1.8; \sigma[Pseudomonas] = 1.48\)). This observation suggests that codon usages are characteristic for a genus rather than for single species. Moreover, there is no correlation between \(\mu\) or \(\sigma\) and the number of reference genes available. This may indicate that reference data sets consisting of 20-30 genes are sufficiently large to get a representative estimate of the mean codon usage.

**Results**

We have analyzed the codon usage of the genes for the restriction endonucleases and the methyltransferases of various type II RM systems to find out whether horizontal gene transfer had contributed to the distribution and evolution of these systems; 29 RM systems could be investigated, because the DNA sequences of their genes are published and reference codon usage data for the hosts are available in the EMBL data bank (Table 1). The results are compiled in Table 2. In six cases the codon usage of the genes for the restriction endonucleases shows \(Z\) scores greater than 2 (EcoRI, EcoRV, KpnI, SphI, SmaI, and ThrHB81). All other scores are uniformly distributed between 1.1 and -1.2. This result is most easily explained by assuming that horizontal gene transfer of genes for restriction endonucleases had occurred in at least 6 of 29 RM systems.

Horizontal gene transfer between two species having a similar codon usage is not detectable by the procedure employed here. To get an estimate of which fraction of horizontal gene transfer events is detectable, we have calculated whether the average codon usages of all 41 species significantly deviate from each other. It turned out that only 436 of 1,640 pairwise comparisons show a significant deviation (\(Z\) score greater than 2). This means that one can estimate that only 27% of all horizontal gene transfer events are detectable. As we found evidence for horizontal gene transfer in 6 of 29 genes for RM systems, it is not unlikely that horizontal gene transfer had occurred in nearly every RM system known today.

As shown in Table 2, the codon usage deviations of