Correlations between genomic GC levels and optimal growth temperatures in prokaryotes

Héctor Musto[a,b], Hugo Naya[a], Alejandro Zavala[a,b], Héctor Romero[a,c], Fernando Alvarez-Valín[b,d], Giorgio Bernardi[b,*]

[a] *Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias, Montevideo, Uruguay*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Naples, Italy*
[c] *Escuela Universitaria de Tecnología Médica, Facultad de Medicina, Montevideo, Uruguay*
[d] *Sección Biomatemáticas, Facultad de Ciencias, Montevideo, Uruguay*

Received 3 June 2004; revised 20 July 2004; accepted 22 July 2004

Available online 30 July 2004

Edited by Takashi Gojobori

Abstract In prokaryotes, GC levels range from 25% to 75%, and T_{opt} from ≈0 °C to >100 °C. When all species are considered together, no correlation is found between the two variables. Correlations are found, however, when Families of prokaryotes are analysed. Indeed, when Families comprising at least 10 species were studied (a set of 20 Families), positive correlations are found for 15 of them. Furthermore, a comparative analysis by independent contrasts made within the Families in order to control for phylogenetic non-independence showed qualitatively equivalent results. We conclude that T_{opt} is one of the factors that influences genomic GC in prokaryotes.

Keywords: Genome evolution; Isochores; DNA thermodynamic stability

1. Introduction

The genomes of prokaryotes cover a broad compositional range, GC levels being approximately comprised between 25% and 75% [1,2]. It was proposed that such range was due to a mutational bias [3,4], a point apparently confirmed by the observation that mutator mutants of *Escherichia coli* could shift their GC levels [5]. Differences in GC levels were also detected between the genomes of cold- and warm-blooded vertebrates [6], previously thought to be very close in base composition. The higher GC levels of the genomes from the latter were explained as due to selection for thermodynamic stability required by DNA, RNA and proteins at their higher body temperatures [7].

At this point, two opposite explanations were competing in order to explain compositional differences in genomes: the neutralist explanation [3,4] and the selectionist explanation [7].

The neutralist explanation was weakened by three observations: (i) that higher GC levels were found in bacteria exposed to UV light, this increase reducing the presence of TpT dinucleotides that could form dimers causing DNA replication problems [8]; (ii) that in some cases high GC levels of prokaryotes were associated with high growth temperatures [9–11] and with aerobiosis [12]; and (iii) that the compositional shifts associated with mutator mutations [5] were within the error limits of the ultracentrifugation approach used to detect them and only concerned "hot spots" [13]. Finally, when a pair of completely sequenced closely related bacteria (*Corynebacterium efficiens* and *C. glutamicum*) were compared, the positive relationship between T_{opt} and GC level was striking [14].

The selectionist hypothesis was (i) apparently weakened by the observation that hyperthermophiles exhibited low GC levels, a point later understood, however, as due to the avoidance of C which can be deaminated at high temperatures [15], and (ii) apparently disposed off by the investigations of Galtier and Lobry [16] and Hurst and Merchant [17] who reported no correlation between T_{opt} and GC (or GC_3). This was interpreted not only as a strong evidence against the thermodynamic hypothesis in prokaryotes, but also as a disproval of the same hypothesis in vertebrates [17–19].

It is important to note, however, that the papers that addressed the thermal stability hypothesis in prokaryotes [16,17] were criticised [20] because many factors, like those quoted above, have often contrasting inputs on genome composition of such a vast array of organisms as prokaryotes, which have been diverging for at least 3.5 billion years [21].

To avoid these drawbacks, we restricted our study of covariation between T_{opt} and genomic GC to the Family level for the following reasons. First, the phylogenetic relationships among prokaryotic Families are still uncertain in several cases (for a review see [22]). In contrast, the phylogenetic relationships among species within each Family are expected to be more accurate, since the times of divergence are much smaller. Second, any method aiming to estimate the correlation between T_{opt} and GC taking into account the phylogenetic relationships needs to infer the character states (T_{opt} and GC) in the internal (ancestral) nodes. Obviously, this inference would be safer for shorter times of divergence (such as within Families) than those inferences that involve nodes connecting different Families. Indeed, as we show here, the failure to detect

* Corresponding author. Fax: +39-81-7641355.
E-mail address: bernardi@szn.it (G. Bernardi).

Abbreviations: GC, guanine plus cytosine; T_{opt}, optimal growth temperature; GC_3, the GC levels of third codon positions

0014-5793/$22.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2004.07.056

correlations when phylogeny was taken into consideration [17] is very likely due to inaccurate inferences in deep internal nodes, which are very probably responsible for hiding the correlation at lower phylogenetic levels.

2. Materials and methods

2.1. Organisms and T_{opt}

Genomic GC and T_{opt} values were taken from [16,23] and from the literature. T_{opt} values were complemented with data collected from the German Collection of Microorganisms and Cell Cultures (http://www.dsmz.de/species/strains.htm). The taxonomic classification was taken from *Taxonomic Outline of the Prokaryotic Genera Bergey's Manual of Systematic Bacteriology*, 2nd Edition, downloaded from http://www.cme.msu.edu/bergeys/.

2.2. Ribosomal sequences

Release 8.1 of the Ribosomal Database Project II [24] provides aligned small subunit ribosomal RNA data for prokaryotes. The trees within each Family were constructed using weighbor software [25] and the matrix generated by DNAdist from the PHYLIP 3.5 package [26].

2.3. Independent contrasts

We used the method of independent contrasts [27] as implemented in the COMPARE 4.5 Software Package [28] to account for phylogenetic non-independence. By taking independent contrasts between species/nodes, we can analyse whether the degree of the difference in T_{opt} between two species/nodes is reflected in a difference of analogous relative level in GC. Then, a regression through the origin was performed.

2.4. Data set

Only Families comprising at least 10 species, and $\Delta T_{opt} > 5$ °C and $\Delta GC > 5\%$, were considered. With these restrictions, 20 Families (that include Bacteria and Archaea) were studied (Table 1). The data are available at http://oeg.fcien.edu.uy/Temperature/.

2.5. Assessing the statistical significance when several correlation coefficients are considered simultaneously

We calculated the correlation coefficients between GC level and T_{opt} in prokaryotic Families. Therefore, under the null hypothesis that T_{opt} and GC are not correlated, we would expect several correlation coefficients to be statistically significant, by chance alone. Specifically, in our sample of 20 Families, we can expect to obtain only one correlation coefficient (positive or negative) to be significant at the 5% level (0.05×20), of which we can expect 0.8 to be significant only at the 5% level [(0.05–0.01) × 20], 0.18 significant only at the 1% level but not at 0.1% level [(0.01–0.001) × 20] and 0.02 significant at the 0.1% level (0.001×20). Therefore, we need to know if, in a set of observed correlations (i.e., in a set of correlation coefficients, each having its own P value), the number of Families that display significant correlations exceeds random expectation by a significant amount. To know this, we followed the approach described in [29], using the multinomial distribution to calculate the probability that by chance alone we could obtain results that are as far, or farther, from random expectation than our results.

3. Results and discussion

When T_{opt} is plotted against genomic GC for the 368 species belonging to the 20 Families, no trend can be detected, although there is, if any, a negative correlation between the two variables (not shown). This result is nearly identical to that reported by Galtier and Lobry [16], although these authors worked at the Genus level.

Table 1 shows the results from the 20 Families studied. Among these taxa, there are 15 which display positive trends (GC increments with T_{opt}), eight of these exhibiting correlation coefficients that are statistically significant. On the other hand, five Families display a negative trend, but only one shows a statistically significant correlation coefficient. The probability of obtaining such distribution of correlation coefficients by chance alone is 4.39×10^{-8}. Importantly, when the analysis is extended to include those Families that have at least five members, the results remain qualitatively the same.

Two conclusions can be drawn from Table 1. First, for each Family the range of variation (Δ) for both T_{opt} and GC is very different. For example, ΔGC varies from 38% (*Enterobacteriaceae*) to 5.5% (*Staphylococcaceae*). ΔT_{opt}, on the other hand, varies from 43.5 °C (*Clostridiaceae*) to 5.5 °C (*Acidamino-*

Table 1
Correlations between T_{opt} and genomic GC within 20 prokaryotic Families

Family	N1	C.c. 1	Significance	N2	C.c. 2	Significance	ΔT	ΔGC
Acetobacteraceae	14	+0.34	NS	7	+0.32	Ns	8.5	14.1
Acidaminococcaceae	11	+0.77	**	7	+0.43	Ns	5.5	22.0
Bacillaceae	18	+0.80	****	13	+0.54	*	50.0	34.5
Chromatiaceae	12	+0.21	NS	7	+0.06	Ns	10.0	23.4
Clostridiaceae	59	+0.20	NS	52	+0.06	Ns	43.5	30.5
Comamonadaceae	22	+0.02	NS	15	+0.24	Ns	16.5	13.5
Corynebacteriaceae	11	−0.67	*	8	−0.37	Ns	13.5	17.3
Enterobacteriaceae	38	+0.54	***	31	+0.13	Ns	15.0	38.0
Eubacteriaceae	11	−0.21	NS	10	+0.11	Ns	7.0	17.0
Flavobacteriaceae	15	−0.02	NS	10	−0.06	Ns	24.0	12.0
Flexibacteriaceae	10	+0.75	*	8	+0.64	+	13.0	15.5
Halobacteriaceae	14	+0.67	**	12	+0.90	****	16.5	9.9
Methanobacteriaceae	12	+0.57	*	6	+0.80	*	28.0	35.2
Microbacteriaceae	15	+0.37	NS	13	+0.23	Ns	6.5	6.8
Micrococcaceae	25	+0.41	*	20	+0.33	Ns	19.5	19.8
Neisseriaceae	23	−0.38	NS	17	+0.05	Ns	12.0	22.5
Pseudomonadaceae	13	+0.63	*	9	+0.62	+	11.0	9.9
Rhodobacteraceae	15	+0.15	NS	14	+0.35	Ns	15.0	14.1
Spirochaetaceae	13	−0.49	NS	11	−0.36	Ns	14.5	37.5
Staphylococcaceae	17	+0.46	+	16	+0.49	*	7.0	5.5

N1, C.c. 1 and N2, C.c. 2 are the numbers of species analysed within each Family and the product–moment (Pearson) correlation coefficients, respectively. In the latter case, the correlations were calculated taking into account the phylogenetic relationships (independent contrasts). Significances are as follows: NS, not significant; *, **, *** and **** are significant at the 5%, 1%, 0.1% and 0.01% levels, respectively. + indicates those coefficients that are at the limit of significance ($0.05 < P < 0.06$). ΔT and ΔGC represent the variation in T_{opt} and genomic GC for each Family.

coccaceae). Although some errors in the taxonomic position of certain species cannot be excluded, this variability is probably related with the time of divergence and rate of change: species which diverged from their last common ancestor more recently, and/or evolve more 'slowly', are expected to share more features, namely the ecological niche, T_{opt}, physiology, etc. This is supported by the correlation found between ΔGC and ΔT_{opt}: $R = 0.52$; $P = 0.02$.

Second, and more important, within most Families there is a link between T_{opt} and GC, and in the majority of cases the correlation coefficient increases (significantly in several cases) with T_{opt}. To sum up, we found that in 15/20 of prokaryotic Families the two variables are positively correlated (eight of them with $P \leqslant 0.05$). Three examples of these correlations are displayed in Fig. 1A–C.

Although these results are clear and suggest that T_{opt} is a factor influencing genomic GC, we cannot rule out the effect of phylogenetic inertia (the fact that closely related species are likely to have similar GC levels), so we used the method of comparative analysis by independent contrasts [27]. This method allows us to see if the GC level shows a correlated response with the adaptation to a new thermal environment. This analysis was carried out on the species belonging to the Families listed in Table 1 for which the 16S RNA were available [24]. Our analysis found that for 17 out of 20 Families there is a positive relation between the two variables, four of them significant and two at the limit of significance. The plots for the same Families shown in Fig. 1 are presented in Fig. 2. On the other hand, for three Families the relation was negative, yet for none of them it was significant. Following the procedure described above for the direct analysis, the probability of getting these results by chance alone is <0.01. The R values within each Family of both analyses (controlling and not controlling for phylogenetic inertia) are significantly correlated ($R = 0.85$, $P < 0.0001$). As can be seen in Table 1, besides the four Families for which there is a significant correlation, there are two more at the limit of significance: *Flexibacteriaceae* and *Pseudomonadaceae*. Since our hypothesis is that GC level not only changes but can increase with T_{opt}, it seems reasonable to apply a one-tailed test. By doing so, we found six Families displaying positive significant correlations coefficients between contrasts, while no negative correlation coefficient was statistically significant (see Table 1). The overall probability of getting by chance this group of correlations with the corresponding significance levels drops to 3.59×10^{-4}.

In addition, when all independent contrasts from different Families (within each taxa) are considered together, they exhibit a positive and significant correlation coefficient (Fig. 3; $R = 0.27$, $P < 0.0001$). Moreover, the increment in T_{opt} was accompanied by an increment in GC in 129 independent contrasts, while 79 contrasts exhibited the opposite behaviour and 76 displayed no changes. If the two parameters were not related, the probability of obtaining this excess of double increments by chance alone is very low ($P < 0.001$, sign test).

In conclusion, we found that T_{opt} and genomic GC are non-independent. In the first place, we have shown that when these two parameters are compared at the Family level they exhibit positive relations in most Families, being statistically significant in several of them. These correlations still hold when the internal phylogenetic relationships are considered. Moreover, when all Families are considered together (but excluding inter-Family comparisons) there is again a significant positive correlation between T_{opt} and GC. We would like to stress that the

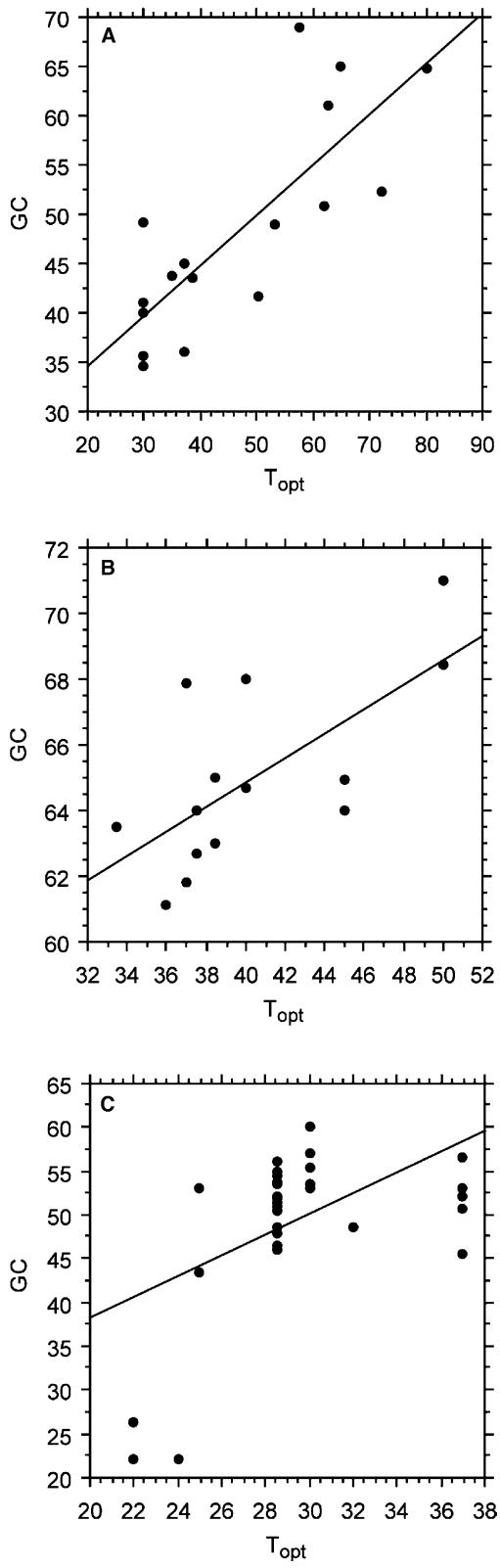

Fig. 1. Plots of T_{opt} vs. genomic GC for *Bacillaceae* (A), *Halobacteriaceae* (B) and *Enterobacteriaceae* (C).

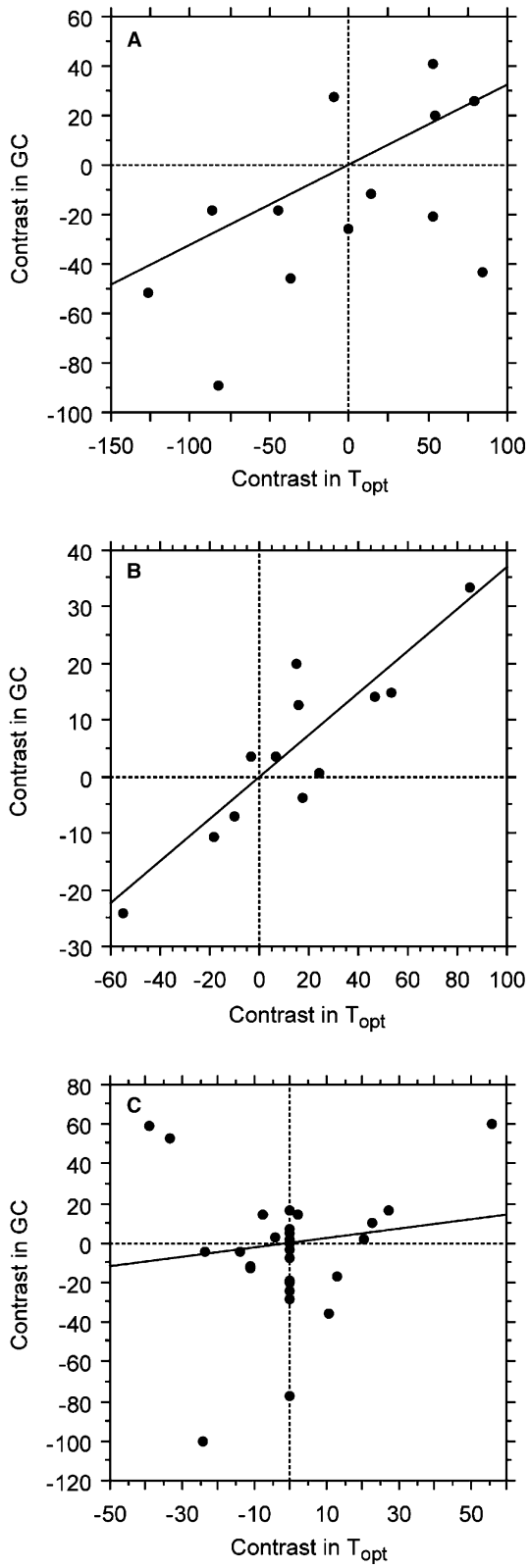

Fig. 2. Contrasts in genomic GC as a function of contrasts in T_{opt} from analyses of *Bacillaceae* (A), *Halobacteriaceae* (B) and *Enterobacteriaceae* (C).

positive correlation becomes evident only when inter-Families comparisons (that are less accurate from many points of view, see Section 1) are excluded from the analysis. It is also safe to

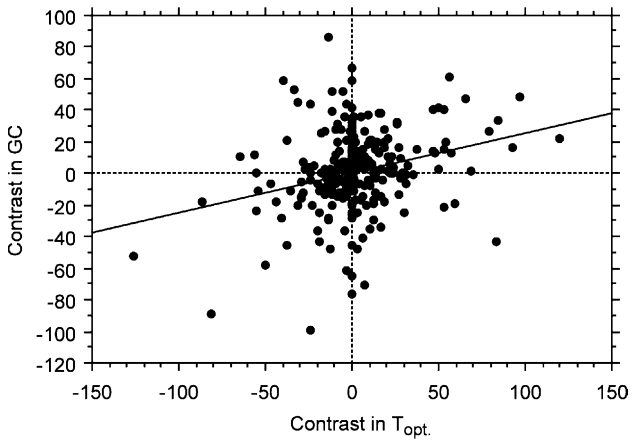

Fig. 3. Plot of contrasts in genomic GC vs. contrasts in T_{opt} for all Families considered.

suppose that when the intra-Families comparison is performed, many variables that could affect the GC level are likely to be more similar.

Finally, we should remark that these results show not only the influence of T_{opt} on genomic GC in prokaryotes, but also that it is not the only one influencing genome composition, as expected from other investigations [8,12,30,31]. Only when a factor becomes predominant, its effect on GC can be clearly seen. Needless to say, the results obtained in these investigations strongly support the idea that base composition is under selection in prokaryotes.

Acknowledgements: This work was partially supported by award 7094 from 'Fondo Clemente Estable', Uruguay.

References

[1] Lee, K.Y., Wahl, R. and Barbu, E. (1956) Ann. Inst. Pasteur 91, 212–224.
[2] Belozerski, A.N. and Spirin, A.S. (1958) Nature 182, 111–112.
[3] Freese, E. (1962) J. Theor. Biol. 3, 82–101.
[4] Sueoka, N. (1962) Proc. Natl. Acad. Sci. USA 48, 582–592.
[5] Cox, E.C. and Yanofsky, C. (1967) Proc. Natl. Acad. Sci. USA 58, 1895–1902.
[6] Thiery, J.P., Macaya, G. and Bernardi, G. (1976) J. Mol. Biol. 108, 219–235.
[7] Bernardi, G. and Bernardi, G. (1986) J. Mol. Evol. 24, 1–11.
[8] Singer, C.E. and Ames, B.N. (1970) Science 170, 822–825.
[9] Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yasuhara, T., Tanaka, T. and Oshima, T. (1984) J. Biol. Chem. 259, 2956–2960.
[10] Claus, D. and Berkeley, R.C. (1986) in: Bergey's Manual of Systematic Bacteriology (Sneath, P.H., Ed.), vol. 2, pp. 1105–1139, Williams and Wilkins, Baltimore.
[11] Whitman, W., Bowen, T. and Boone, D. (1992) in: The Prokaryotes (Balows, A., Trueper, H., Dworkin, D., Harder, W. and Schleifer, K., Eds.), pp. 719–767, Springer–Verlag, New York.
[12] Naya, H., Romero, H., Zavala, A., Alvarez, B. and Musto, H. (2002) J. Mol. Evol. 55, 260–264.
[13] Bernardi, G. (1993) Mol. Biol. Evol. 10, 186–204.
[14] Nishio, Y., Nakamura, Y., Kawarabayasi, Y., Usuda, Y., Kimura, E., Sugimoto, S., Matsui, K., Yamagishi, A., Kikuchi, H., Ikeo, K. and Gojobori, T. (2003) Genome Res. 13, 1572–1579.
[15] Ehrlich, M., Norris, K.F., Wang, R.Y., Kuo, K.C. and Gehrke, C.W. (1986) Biosci. Rep. 6, 387–393.
[16] Galtier, N. and Lobry, J.R. (1997) J. Mol. Evol. 44, 632–636.
[17] Hurst, L.D. and Merchant, A.R. (2001) Proc. R. Soc. Lond. B. Biol. Sci. 268, 493–497.

[18] Ream, R.A., Johns, G.C. and Somero, G.N. (2003) Mol. Biol. Evol. 20, 105–110.
[19] Gautier, C. (2000) Curr. Opin. Genet. Dev. 10, 656–661.
[20] Bernardi, G. (2004) Structural and Evolutionary Genomics. Natural Selection in Genome Evolution. Elsevier, Amsterdam.
[21] Hedges, S.B. (2002) Nat. Rev. Genet. 3, 838–849.
[22] Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V. (2002) Trends Genet. 18, 472–479.
[23] Holt, J.G., Krieg, N.R., Sneath, P.H.A., Staley, J.T. and Williams, S.T. (1994) Bergey's Manual of Determinative Bacteriology. William and Wilkins, Baltimore.
[24] Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., McGarrell, D.M., Schmidt, T.M., Garrity, G.M. and Tiedje, J.M. (2003) Nucleic Acids Res. 31, 442–443.
[25] Bruno, W.J., Socci, N.D. and Halpern, A.L. (2000) Mol. Biol. Evol. 17, 189–197.
[26] Felsenstein, J. (1993) Distributed by the author. Department of Genetics, University of Washington, Seattle. Available from: http://evolution.gs.washington.edu/phylip/.
[27] Felsenstein, J. (1985) Amer. Naturalist 125, 1–15.
[28] Martins, E. P. (2003) Distributed by the author, Department of Biology, Indiana University, Bloomington, IN. Available from: http://compare.bio.indiana.edu/.
[29] Alvarez-Valin, F., Jabbari, K., Carels, N. and Bernardi, G. (1999) J. Mol. Evol. 49, 330–342.
[30] McEwan, C.E., Gatherer, D. and McEwan, N.R. (1998) Hereditas 128, 173–178.
[31] Rocha, E.P. and Danchin, A. (2002) Trends Genet. 18, 291–294.