

Evolutionary and comparative genomics of *Leptospira*

DOI: 10.3395/reciis.v1i2.Sup.103en



Natalia Rego

Unidad de Bioinformática,
Institut Pasteur de Montevideo,
Montevideo, Uruguay
nrego@pasteur.edu.uy



Hugo Naya

Unidad de Bioinformática,
Institut Pasteur de Montevideo,
Montevideo, Uruguay
naya@pasteur.edu.uy

Guillermo Lamolle

Unidad de Bioinformática, Institut Pasteur de Montevideo / Sección Biomatemática, Facultad de Ciencias, Universidad de la República (UdelaR), Montevideo, Uruguay
glamolle@pasteur.edu.uy

Fernando Álvarez-Valin

Sección Biomatemática, Facultad de Ciencias, Universidad de la República (UdelaR), Montevideo, Uruguay
falvarez@fcien.edu.uy

Abstract

The availability of complete genome sequences has allowed the study of genic and nucleotide content changes during the evolution of pathogenic and symbiont bacterial lineages. Here we present the comparative genomics between four strains belonging to two species, *Leptospira interrogans* (*L. interrogans* Lai str. 56601 and *L. interrogans* Copenhageni str. Fiocruz L1-130) and *L. borgpetersenii* (*L. borgpetersenii* serovar Hardjo-bovis str. JB197 and *L. borgpetersenii* serovar Hardjo-bovis str. L550). Strains from the latter species have reduced their host range being restricted to a host-to-host transmission cycle while the former species could survive for long periods in fresh water. A detailed compositional and substitutional analysis of 2416 tetrads of orthologs indicates that this change in environmental condition is accompanied by changes in genomic GC content that affected the entire genome. Moreover, this analysis reveals that the interspecies divergence is surprisingly large, with a synonymous/amino acid distances ratio equal to 7.17, a ratio much larger than in other groups. We show that this increased ratio is an indication of acceleration in substitutions rates that especially affected synonymous positions and is in connection with the change in base composition already described. We hypothesize that this process specifically took place in the branch that leads to *L. borgpetersenii*.

Keywords

Leptospira, genome reduction, GC content, substitution rates, host range

Introduction

The order Spirochaetales comprises several species of long and slender bacteria whose monophyletic origin is evident from multiple structural features that are unique to this group (noteworthy the helical protoplasmic cylinder encased by an outer sheath). Molecular phylogenetic reconstruction based on the 16S rRNA are in agreement with these ultrastructural data, despite spirochetal clades branch deeper than clades in other bacterial orders (PASTER et al., 1991, p.6104-6108). The ecological range of spirochetes is very wide, extending from free-living organisms to parasitic forms and from aerobic to anaerobic species.

During its evolutionary history, spirochetes have undergone very significant variations in genome size and composition. One of these changes is depicted in Figure 1, where GC content variability is represented in

a phylogenetic perspective. This figure shows that the variability is not restricted to intergenus comparisons but most strikingly, there are several species belonging to a same genus, and hence expected to be evolutionary tightly related, whose genomic GC contents are remarkable different. As an extreme example, we can mention the genus *Treponema*, where *Treponema pallidum* (causative agent of syphilis) and *Treponema denticola* (periodontal plaque spirochete) exhibit genomic GC contents of 53% and 38% respectively. Another striking example is given by species of the genus *Leptospira*, whose phylogenetic relatedness is even closer, yet they still exhibit very significant differences in genomic GC. At the moment, the spirochetal base composition heterogeneity has been discussed as a diagnostic character for taxonomic purposes (OLSEN et al., 2000, p.45-46) but has not yet been assessed in an evolutionary framework.

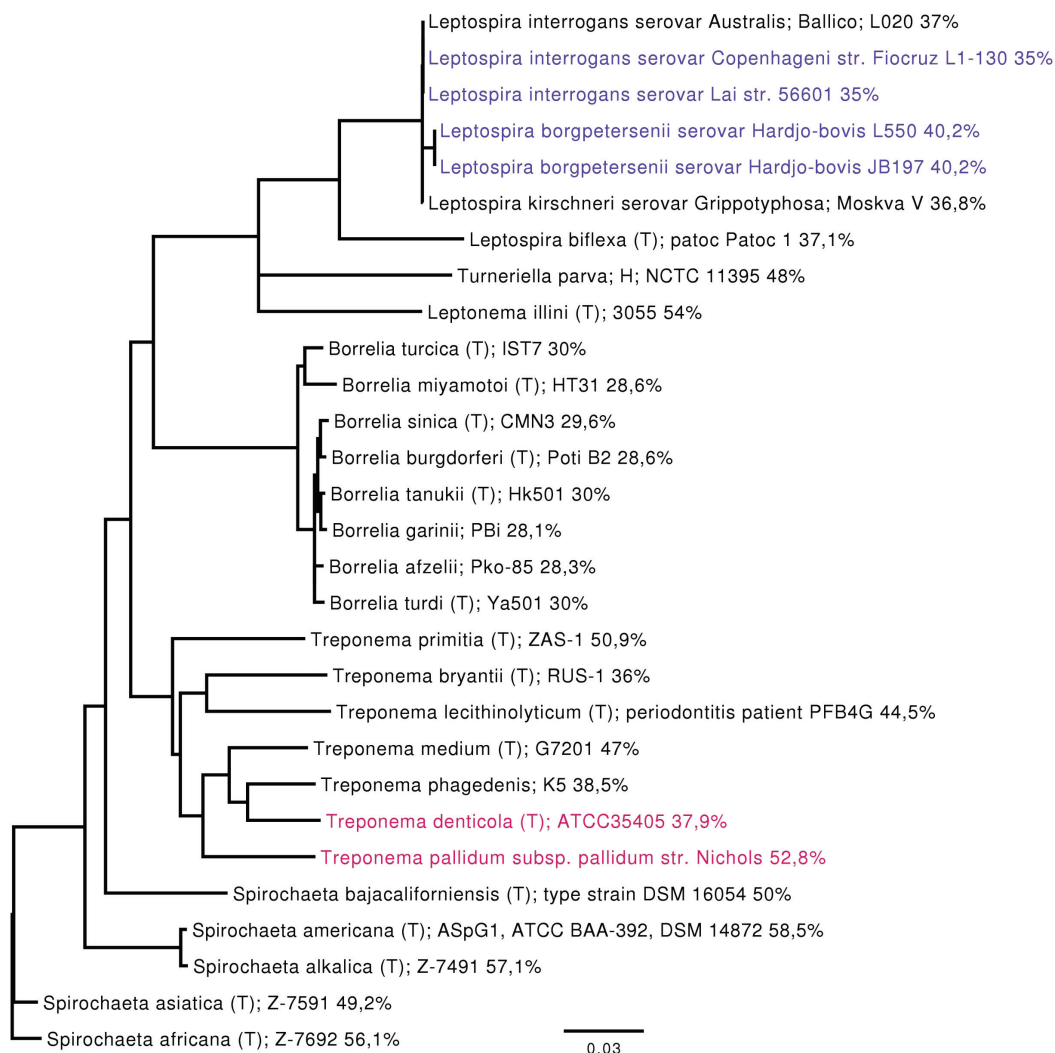


Figure 1 - Phylogenetic relationships among 29 spirochetes inferred from 16S rRNA sequences obtained at The Ribosomal Database Project RDP-II: <http://rdp.cme.msu.edu/>. The tree was built with Weighbor, the method implemented at this site. The genomic GC content for each taxa is shown. Branch lengths are proportional to the number of substitutions. The reference scale (that appears at the bottom) represents 0.03 nucleotide substitutions per site. The species of interest are highlighted with different colours.

Not only in the *Treponema* species mentioned above, but also for the species pair *L. interrogans* – *L. borgpetersenii*, the genome size difference has been explained as a result of genomic expansion in one species (by duplication and horizontal gene transfer) and genomic reduction in the other (e.g. IS mediated gene loss). Both processes, genome reduction and horizontal gene transfer, are related to the different lifestyles found in the genus *Leptospira*. This was proposed based on the following two observations: first, genomic reduction parallels host range decrease, and second, the process of gene loss is not random but mainly affects genes whose presence would allow environmental sensing and wider capability for metabolite transport and utilization (BULACH et al., 2006, p.14562-14564; SESHADRI et al., 2004, p.5648-5649). Unlike other microbial genome reductions, where there exists a genome AT content enrichment associated with host specificity increase (e.g. *Buchnera*, *Mycobacterium*; MORAN, 2002, p.585), these spirochetes show the opposite trend, that is to say, genomic GC content increases in association with higher host specificity.

In this work, we analyze several evolutionary aspects within the genus *Leptospira* with special emphasis on the processes that involve changes in genomic GC and nucleotide substitution rates, trying to characterize in detail these evolutionary processes and to shed some light on the evolutionary forces that could have driven them. For this purpose, we concentrate on four strains whose genomes have been completely determined: two strains belonging to *L. interrogans* (*L.interrogans* Lai str. 56601 and *L. interrogans* Copenhageni str. Fiocruz L1-130) and two strains from *L. borgpetersenii* serovar Hardjo-bovis (str. JB197 and str. L550). On average, the latter are 17% smaller in genome size and 5% higher in GC content.

Both species have been extensively studied because they are the causative agents of most cases of leptospirosis, one of the most widespread zoonoses, responsible of more than half a million human cases per year worldwide. Although the two produce clinical symptoms of this disease that are similar, they are epidemiologically very different. While *L. interrogans* reaches wet soil or

freshwater from the host urine and is capable of surviving more than 200 days in aquatic environments until it infects a new mammalian host, *L. borgpetersenii* does not tolerate nutrient deprivation and is limited to a direct host to host transmission cycle (BULACH et al., 2006, p.14562-14563).

Materials and Methods

The complete genome sequences from the strains analyzed here were downloaded from public databases. Orthologs available in the 4 strains were identified using Blastp by best reciprocal hits (ALTSCHUL et al., 1990). By doing this, we were able to identify 2419 groups of orthologous genes that are present in the four strains. Each of these groups was accurately aligned at the amino acid level using MUSCLE (EDGAR, 2004) and then backtranslated into the known DNA sequence.

Synonymous and non-synonymous distances were calculated using the method of NEI and GOJOBORI (1985) with the modifications suggested by ZHANG et al. (1998) that correct for transition/transversion biases which mainly affect the way of counting the number of synonymous and non-synonymous sites in the third codon positions of duets (two-fold degenerated codons). In this work, the correction was done according to the transition/transversion ratio observed at the third codon positions of quartets.

Comparison of compositional patterns in orthologous genes

A detailed compositional and substitutional analysis of the 2416 tetrads of orthologs was carried out and is presented in Table 1. It should be stressed that all calculations were performed after eliminating the alignment gaps, thus guaranteeing that the segments compared for each gene are strictly homologous. This ensures that any differences that could be detected among the strains in the comparison cannot be attributed to differential deletions or insertions taking place in a given strain, but exclusively to the nucleotide substitution process.

Table 1 - Compositional and substitutional analysis of the 2416 tetrads of orthologs

Serovar-Strain	GC1	GC2	GC3
<i>L. borgpetersenii</i> Hardjo-bovis str. JB197	0.4772	0.3520	0.3933
<i>L. borgpetersenii</i> Hardjo-bovis str. L550	0.4772	0.3521	0.3933
<i>L. interrogans</i> Lai str. 56601	0.4600	0.3453	0.2813
<i>L. interrogans</i> Copenhageni str. Fiocruz L1-130	0.4601	0.3453	0.2812
T-test for paired comparisons	37.5, P<<10-12	18.2, P<<10-12	112, P=0

The first evident observation that can be drawn from Table 1 is that both *L. interrogans* strains and both *L. borgpetersenii* strains are very similar to each other, yet they are consistently different when the comparison is done between strains belonging to different species. In effect, both *L. borgpetersenii* strains are GC-richer than both *L. interrogans* strains, being these differences consistent and statistically very highly significant (*t*-test for paired comparisons) for the three codon positions. Nevertheless, the differentiation is especially pronounced for the third codon position where the average difference is larger than 10%.

It becomes necessary to verify that this behavior is not a simple and spurious average differentiation, caused by some genes with different GC content acquired perhaps by horizontal gene transfer, but that it

affects the complete genomes and the majority of their constituent genes. For this purpose, we plotted the GC level of one species against the other for each gene and for each codon position (Figure 2). The diagonal line was drawn in these figures to indicate the place around which the points (genes) should symmetrically fall if the distribution was not biased. In other words, if the GC content of a given gene is the same in both species, then the point corresponding to that gene should fall exactly on the line. However, as is evident from these plots, the vast majority of points are located above the diagonal line, thus indicating that for these genes the GC level is higher in *L. borgpetersenii*. In agreement with the results presented in Table 1, the deviations above the diagonal line are stronger in GC₁ than in GC₂ but they are particularly strong for the third codon position.

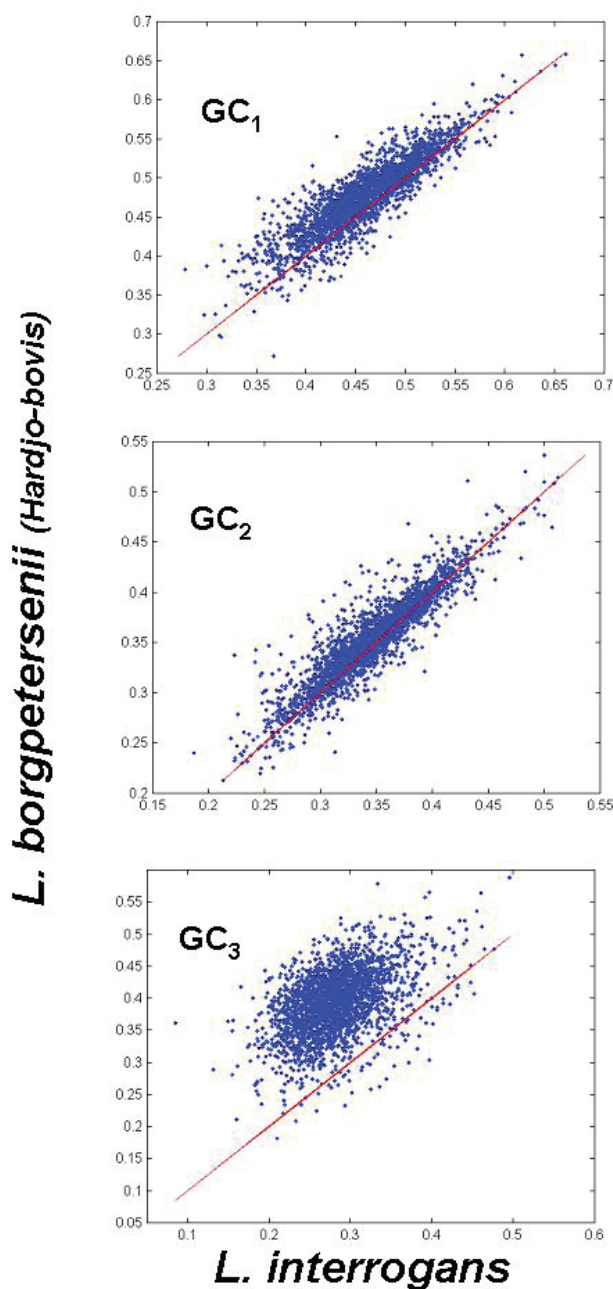


Figure 2 - Scatter plot among GC levels in *Leptospira* for each gene and for each codon position.

Such strong differences in GC composition in the genus *Leptospira* become particularly puzzling because *L. borgpetersinii* strains, which lost the ancestral ability to survive out of the host, are notoriously GC richer than the related *L. interrogans*, in contradiction with the expected behavior for the GC content in genome reductions associated with parasitism (ROCHA; DANCHIN, 2002, p.291-294). On the other hand, the decrease in niche breadth corresponding to the genome reduction in one group of *Leptospira* could lead to an increase in the importance of otherwise relatively minor factors

such as the optimal growth temperature (MUSTO et al., 2004).

Evolutionary rates in *Leptospira*: acceleration of silent rates

We estimated the amino acid and synonymous distances for the 2419 groups of orthologs. The results of these analyses are presented in Table 2 and Figure 3. The latter shows the frequency distribution of intergroup comparisons, for both synonymous and amino acid distances, panels a and b respectively.

Table 2 - Nucleotide distances in *Leptospira*

Serovar-Strain	1	2	3	4
Synonymous rates				
1- <i>L. interrogans</i> Copenhageni str. Fiocruz L1-130	0			
2- <i>L. interrogans</i> Lai str. 56601	0.0038	0		
3- <i>L. borgpetersenii</i> Hardjo-bovis str. L550	1.2266	1.2273	0	
4- <i>L. borgpetersenii</i> Hardjo-bovis str. JB197	1.2257	1.2253	0.0115	0
Amino acid rates				
1- <i>L. interrogans</i> Copenhageni str. Fiocruz L1-130	0			
2- <i>L. Interrogans</i> Lai str. 56601	0.0021	0		
3- <i>L. borgpetersenii</i> Hardjo-bovis str. L550	0.1687	0.1687	0	
4- <i>L. borgpetersenii</i> Hardjo-bovis str. JB197	0.1688	0.1687	0.003	0
Small subunit rRNA genes				
1- <i>L. interrogans</i> Copenhageni str. Fiocruz L1-130	0			
2- <i>L. interrogans</i> Lai str. 56601	0.00072	0		
3- <i>L. borgpetersenii</i> Hardjo-bovis str. L550	0.007889	0.007889	0	
4- <i>L. borgpetersenii</i> Hardjo-bovis str. JB197	0.007883	0.007883	0	0

These results deserve some comments. In the first place the distances inside each subgroup (species) are very small, both at the synonymous and amino acid level. On the other hand, the intergroup distances are quite large, on average around 0.17 amino acid changes per codon (for amino acid changes), while silent divergence attains 1.22 changes per synonymous site (on average) which is in the range of saturation. Such an extent of silent nucleotide divergence is startling if one takes into account the fact that species belonging to the same genus are compared. Finally and more importantly, the ratio

synonymous/amino acid distances is 7.17 (1.22/0.17), a remarkable large figure.

In order to compare the amount of divergence (as well as their ratio) observed in *Leptospira*, it becomes necessary to have a reference group for which information on divergence times and substitution rates is available. These distances were thus compared to the amount of divergence among different mammalian orders. It is worth reminding that mammalian orders diverged from each other around 100 million years ago (Mya) (SPRINGER et al., 2003). Interestingly enough,

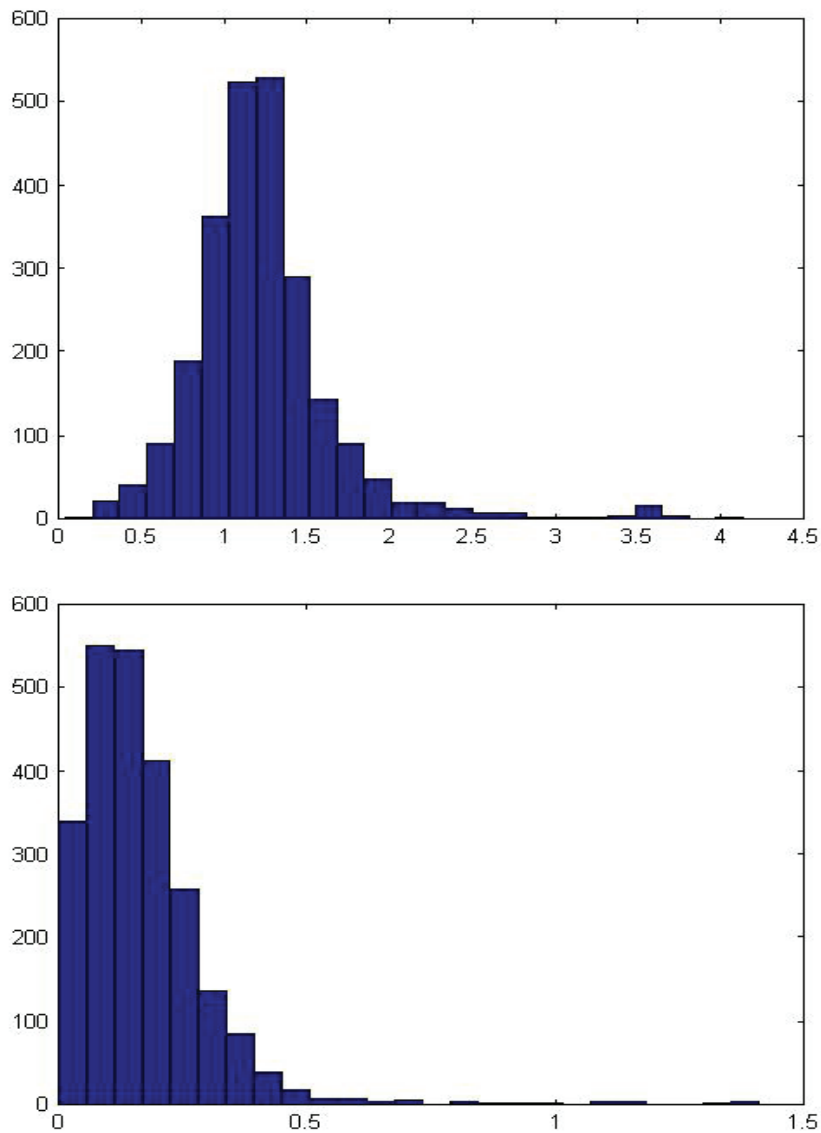


Figure 3 - Distribution of (a) synonymous and (b) amino acid distances (Poisson correction) between orthologous genes from *L. interrogans* and *L. borgpetersenii* species.

for mammals the synonymous distances are on average 0.35 changes per site (this average was obtained from 50 different alignments of orthologous genes comprising 4 different mammalian orders, taken from Alvarez-Valin et al., 1998), which is less than one third of the average synonymous distance observed between *L. interrogans* and *L. borgpetersenii* species. In contrast, the average amino acid distance among mammalian orders is 0.21 changes per codon, notoriously larger than that between *L. interrogans* and *L. borgpetersenii*. Thus, the ratio synonymous/amino acid distances in mammals is 1.67, one quarter of that among leptospiras.

We also compared the amount of nucleotide distance in *Leptospira* versus mammals for small subunit rRNA genes (namely 16S rRNA in *Leptospira* and 18S rRNA in mammals). The picture that emerges from this comparison

is again the inverse from that exhibited by synonymous distances, that is, the distances between *L. interrogans* and *L. borgpetersenii* (0.79%; Table 2) are shorter than those between mammalian orders (1.1%; not shown).

At first glance, this disparity in the behavior between silent positions on the one hand, and amino acid coding positions (or ribosomal genes) on the other, seems somewhat puzzling. However, it is worth taking into account that the evolutionary factors that tend to increase the nucleotide substitution rates (e.g. generation time, mutational rate, etc.) are expected to produce a larger effect on synonymous rates than on non-synonymous ones (OHTA, 1995) or in ribosomal genes. In other words, the acceleration in substitution rates most likely affected the whole genome, but appeared particularly strong at synonymous positions.

We hypothesize that this acceleration has a connection with the change in base composition already described, and took place specifically in the branch that leads to the *L. borgpetersenii* group. In the next section we present evidence that gives support to this hypothesis.

Ancestral composition in *Leptospira*

As was described above, *Leptospira interrogans* and *L. borgpetersenii* differ in their genomic GC content, this difference being particularly strong at the third position of codons. Three possible evolutionary scenarios could be envisaged. The first one is that *L. interrogans* remains in the ancestral condition while *L. borgpetersenii* underwent an increase in genomic GC content. Alternatively, it can be postulated that *L. borgpetersenii* represents the ancestral condition and hence *L. interrogans* was subjected to a decrease in genomic GC content. Finally, another yet less parsimonious scenario is that none represents the ancestral condition.

In order to be able to discern the direction of the change in base composition, it becomes necessary to determine the ancestral condition. As it emerges from Figure 1, *Leptospira biflexa* branches off much earlier than *L. interrogans* and *L. borgpetersenii*, it thus can be used as an outgroup to infer the ancestral GC composition inside the *Leptospira* genus. A visual inspection of this figure immediately suggests that the first scenario is the most probable one, that is to say, *L. interrogans* represents the ancestral condition, since *L. biflexa* (37% GC) and *L. interrogans* (between 35 and 37% GC) have very similar genomic GC content, notoriously lower than that of *L. borgpetersenii* (40.2 %).

Conclusions

The genus *Leptospira* is particularly interesting because strains from *L. borgpetersenii*, which lost the ancestral ability to survive out of the host, are notoriously GC richer than the close relative *L. interrogans* group. This observation is in clear contradiction with what would be the expected behavior for the GC content in genome reductions associated with parasitism. From the comparison of orthologous genes, we have shown that differences in GC content between both groups pervade the entire genome, even in the non-synonymous codon positions, but it is particularly strong for synonymous positions.

We also analyzed the ratio between synonymous and non-synonymous distances between groups of *Leptospira* and compared these results with the ratios obtained from different mammalian orders that diverged from each other around 100 Mya. Surprisingly, the synonymous distance for mammalian orders is less than one third of that corresponding to *Leptospira* groups. By contrast, the average amino acid distance among mammalian orders is notoriously larger than that between *L. interrogans* and *L. borgpetersenii* groups. These opposite trends in nucleotide distances lead to strong differences in the ratio of synonymous to non-synonymous substitution rates, and

testify of a genome-wide acceleration in substitution rates that specially affected the third position of codons in protein coding genes.

Due to the scarcity of information on sequence data from other species of this genus, it is relatively difficult to faithfully discern its ancestral GC content. However, the fragmentary data available seems to favor the hypothesis of an ancestral GC content similar to that of present day *L. interrogans*.

In summary, the global picture shows remarkable changes in compositional levels and substitution rates among groups from the *Leptospira* genus. It would be interesting to further investigate possible biological and evolutionary forces driving such processes. Specifically, it would be of interest to determine whether the observed change in composition could be attributed either to changes in the underlying pattern of mutations or to functional factors (i.e. selective reasons). The first possibility could be tested by determining the pattern of substitutions in pseudogenes that is widely recognized to reflect the mutational spectrum.

As long as functional reasons are concerned, a possible explanation for the increase in genomic GC in *L. borgpetersenii* could be found in the thermodynamic hypothesis proposed by BERNARDI and BERNARDI (1986) and more recently supported by results from MUSTO et al. (2004; 2005). According to this hypothesis, the increase in GC would be an adaptive response to increasing environmental temperature. This hypothesis fits the results presented here provided that *L. borgpetersenii* is restricted to the mammalian (warm – blooded) host.

Acknowledgments

We are indebted to Rosina Piovani for helpful discussions. This work was partially supported by grant FPTA-252 “Desarrollo de capacidades bioinformáticas en el área de anotación genómica” from Instituto Nacional de Investigación Agropecuaria (Uruguay).

Bibliographic references

- ALTSCHUL, S.F. et al. Basic local alignment search tool. **Journal of Molecular Evolution**, v.215, n.3, p.403-410, 1990.
- ALVAREZ-VALIN, F.; JABBARI, K.; BERNARDI, G. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. **Journal of Molecular Evolution**, n.46, p.37-44, 1998.
- BERNARDI, G.; BERNARDI, G. Compositional constraints and genome evolution. **Journal of Molecular Evolution**, v.24, n.1-2, p.1-11, 1986.
- BULACH, D.M. et al. Genome reduction in *Leptospira borgpetersenii* reflects limited transmission potential. **PNAS**, v.103, n.39, p.14560-14565, 2006. Available at: <www.pnas.org/cgi/doi/10.1073/pnas.0603979103>.
- EDGAR, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. **BMC**

Bioinformatics, v.5, n.113, 2004. [doi:10.1186/1471-2105-5-113].

MORAN, N.A. Microbial minimalism: genome reduction in bacterial pathogens. **Cell**, v.108, p.583-586, 2002.

MUSTO, H. et al. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. **FEBS Letters**, v.573, n.1-3, p.73-77, 2004.

MUSTO, H. et al. The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor. **Biochem Biophys Res Commun**, v.330, p.357-360, 2005. [doi:10.1016/j.bbrc.2005.02.133].

NEI, M.; GOJOBORI, T. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. **Molecular Biology and Evolution**, v.3, n.5, p.418-426, 1986.

OLSEN, I.; PASTER, B.J.; DEWHIRST, F.E. Taxonomy of spirochetes. **Anaerobe**, n.6, p.39-57, 2000. [doi:10.1006/anae.1999.0319].

OHTA, T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory.


Journal of Molecular Evolution, v.40, n.1, p.56-63, 1995.

PASTER, B.J. et al. Phylogenetic analysis of the Spirochetes. **Journal of Bacteriology**, v.173, n.19, p.6101-6109, 1991.

ROCHA, E.P.; DANCHIN, A. Base composition bias might result from competition for metabolic resources. **TRENDS in Genetics**, v.18, n.6, p.291-294, 2002.

SESHADRI, R. et al. Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. **PNAS**, v.101, n.15, p.5646-5651, 2004. Available at: <www.pnas.org/cgi/doi/10.1073/pnas.0307639101>.

SPRINGER, M.S. et al. Placental mammalian diversification and the Cretaceous-Tertiary boundary. **PNAS**, v.100, n.3, p.1056-1061, 2003. Available at: <www.pnas.org/cgi/doi/10.1073/pnas.0334222100>.

ZHANG, J.; ROSENBERG, H.F.; NEI, M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. **PNAS**, v.95, n.7, p.3708-3713, 1998. 

About the authors

Natalia Rego

Received the degree in Biology from Universidad de la República - Uruguay in 2006. She is currently a MSc student in zoology. Her work is on bioinformatics and evolutionary genomics at the Institut Pasteur de Montevideo.

Hugo Naya

Received the degree in Biology in 1999 and the PhD degree in biology (computational genomics) from Universidad de la República - Uruguay in 2004. His work involves computational genomics and quantitative genetics. Currently is the head of the Bioinformatics Unit at the Institut Pasteur de Montevideo.