Universidad Nacional Autónoma de México

Instituto de Biotecnología

Programa de Doctorado en Ciencias Bioquímicas

Análisis de la Curvatura Estática del DNA Genómico.

Tesis

Que para obtener el grado de
Doctor en Ciencias
Presenta:

Ruy Jáuregui Sandoval

B

El presente trabajo fué realizado en el Instituto de Biotecnología de laUNAM como parte de los estudios del programa de Doctorado en Ciencias Bioquímicas, bajo la dirección del Dr. Enrique Merino Pérez, y con el apoyo de la beca con número de registro 138377 del Consejo Nacional para la Ciencia y la Tecnología, El Programa de Apoyo a los Estudios de Posgrado No. 203341, y el Programa de Apoyo a Proyectos de Investigación e Inovación Tecnológica No. IN230798.

Índice:

Introducción.

Desarrollo.

Resumen.

La curvatura del DNA ha sido estudiada por más de 20 años, y se ha relacionado con un amplio espectro de funciones biológicas, tales como la replicación, la regulación transcripcional, el empacamiento del DNA, la integración de virus y transposones, y la recombinación (1, 7, 14, 24). Estos estudios fueron realizados principalmente en regiones discretas y *loci* específicos. No fue sino hasta que el avance en la tecnología de secuenciación permitió obtener la secuencia de genomas bacterianos completos que los determinantes globales de la curvatura de un genoma pudieron ser analizados. En esta tesis se presenta una de las primeras caracterizaciones de los determinantes globales de los perfiles de curvatura de genomas totales, incluyendo la relación que tiene la curvatura del DNA con las preferencias de uso de codones y la composición del proteoma.

Estudios de la relación que hay entre la curvatura del DNA y la regulación transcripcional han sido realizados para un número relativamente pequeño de genes, tales como los genes dependientes de sigma 54 *glnA* y *glnH* en *E. coli* (12), el regulador transcripcional H-NS (23), sigma S (13), y CRP e IHF (14). Recientemente se han publicado estudios de carácter global, en donde se ha reportado la curvatura promedio de las regiones promotoras de *E. coli* y de *H. sapiens* (16), posteriormente se ha publicado una comparación entre los perfiles genómicos promedio de regiones reguladoras de genomas de bacterias mesófilas e hipertermófilas (17), así como las características de la curvatura de promotores de genes de micobacterias (18).

En el presente trabajo extendemos estas nociones previas sobre la trascendencia de la curvatura del DNA como elemento de la regulación transcripcional dentro de un contexto genómico, mediante la evaluación de la conservación de señales de curvatura en regiones de regulación de genes ortólogos en más de 90 genomas. Hemos identificado varios grupos de proteínas ortólogas que presentan señales de curvatura significativamente conservadas, indicando un posible mecanismo común de regulación. Estos grupos son examinados y discutidos en este trabajo.

Nuestros datos nos han permitido identificar grupos de genes corregulados en *E. coli* que presentan señales de curvatura conservadas, en varios casos no hay una descripción previa que relacione a la proteína reguladora con la curvatura del DNA. Este análisis sienta las bases de la caracterización experimental del mecanismo de regulación de varios genes tales como los dependientes de TyrR, en el cual la curvatura del DNA pudiera jugar un papel importante.

Abstract.

DNA curvature has been studied for more than 20 years and has been related to a broad spectrum of biological functions such as DNA replication, transcriptional regulation, DNA packaging, transposon and virus integration, and recombination (reviewed by 24, 7, 1, 14). It was not until the advent of powerful sequencing techniques and the subsequent growth on the DNA sequence databases, along with the determination of several complete genomic sequences, that DNA curvature was able to be studied in a genomic context.

This work describes one of the earliest characterizations of the global determinants of a genome's curvature profile, including its relationship with the codon usage preferences and the proteome composition. Studies on the relationship of DNA curvature with transcriptional regulation have been conducted for a relatively small number of genes and discrete *loci*, such as Sigma-54 dependent *glnAp2* and *glnHp2* genes (12), H-NS histone-like protein (23), Sigma-s (13), IHF and CRP regulatory proteins (14), and artificial constructs using the T7 virus promoter (25).

Within a more global scope analysis Gabrielian and co-workers (16) found that the promoter regions of *E. coli* tend to be significantly more curved than coding regions or randomly permuted sequences. Bolshoy and Nevo (17) reported high average curvature values in the upstream regions of mesophilic bacteria, as opposed to the case of hyperthermophilic bacteria and Archaea. Here we extend the previous notions about the transcendence of DNA curvature in the transcriptional regulation of genes within a genomic scope.

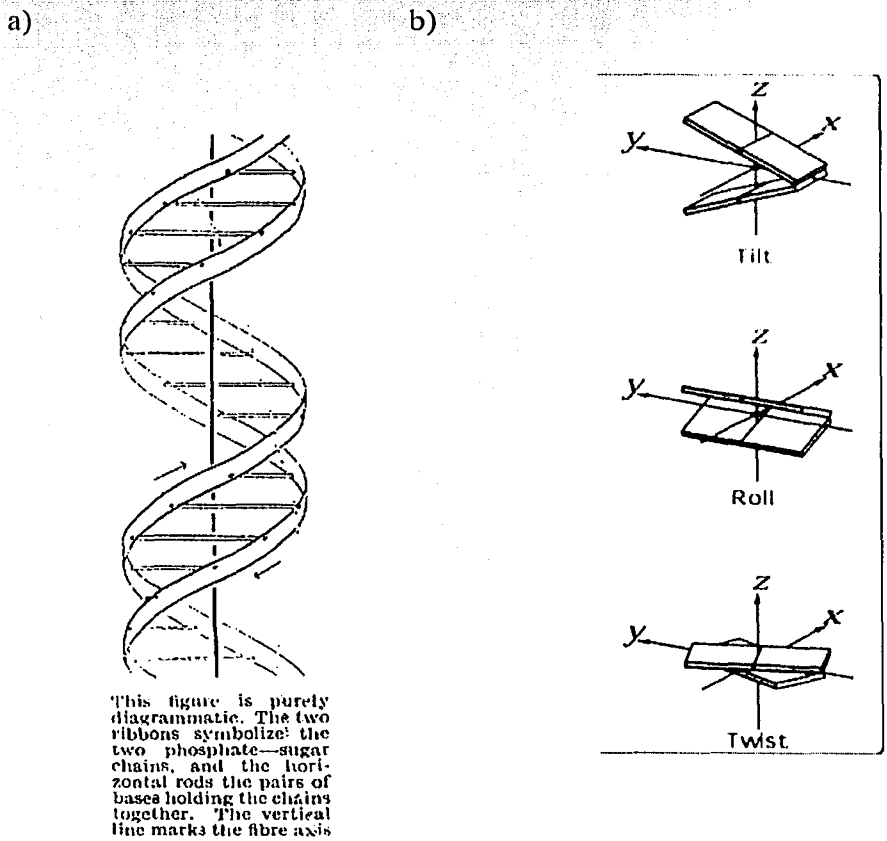
In this study we evaluate the conservation of DNA static curvature as a regulatory element in the transcription initiation of eubacterial and archaeal genes from 90 complete genomes. Significant curvature signals were collected and conserved curvature profiles were identified in orthologous gene sets, as defined in the Cluster of Orthologous Groups (COGs) (26) database and additional homology data for the genomes not yet included. Several orthologous gene sets present significantly conserved curvature signals, indicating a possible common regulatory mechanism. Relevant examples of orthologous groups with conserved curvature signals are examined and discussed.

Our data allowed us to identify several genes corregulated in the *E. coli* genome, with conserved curvature signals. This gives a mainframe for the experimental characterization of regulatory mechanisms related to DNA curvature previously unknown, such as the case of the PurR regulator.

Introducción.

La estructura del DNA.

La estructura de la molécula de DNA, caracterizada inicialmente por Watson y Crick en 1952, fue representada como una doble hélice, siguiendo una trayectoria recta. Estudios posteriores demostraron que las interacciones electromagnéticas entre los diferentes nucleótidos causan variaciones a lo largo de la trayectoria de la hebra de DNA. Estas han sido descritas mediante un conjunto de variables geométricas (1) (fig 1).



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Figura 1. a) Imagen original del articulo de Watson y Crick (Nature, 1953) b) Algunas de las variables geométricas que pueden desviar al trayectoria de la molécula de DNA, la variable twist presenta un valor constante de 36 grados, las otras variables pueden generar cambios en la trayectoria.

Las variables tilt y roll, nos permiten reconstruir la trayectoria de la molécula de DNA en el espacio. Esta trayectoria depende intrínsicamente de la secuencia, y la representación gráfica de la misma nos permite identificar y visualizar regiones curvas en la molécula del DNA (fig 2).

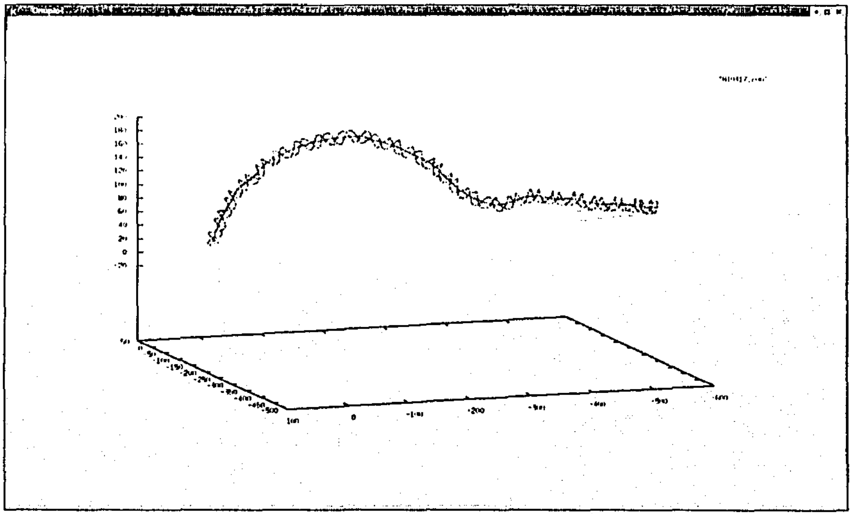


Figura 2. Mapeo de la trayectoria de un fragmento de DNA curvo.

El DNA curvo.

Las primeras nociones acerca del DNA curvo provienen de la observación de patrones de corrimiento anómalo de fragmentos de DNA en procesos de electroforesis en geles de agarosa (2). La caracterización de estos fragmentos de DNA reveló que las restricciones estructurales impuestas por una región curva en la molécula no permite el libre flujo a través de la retícula del gel, causando una disminución en la velocidad de migración de la misma.

Originalmente fueron propuestos dos modelos para explicar el fenómeno del DNA curvo: El modelo del vecino cercano (3) propone que las regiones curvas en el DNA son producto de la acumulación de muchas pequeñas deformaciones locales en la molécula. El modelo de unión (4) propone que los sitios curvos ocurren en la interfase entre dos variantes estructurales diferentes de la doble hélice de B-DNA. Posteriormente, estos modelos han derivado en diferentes sub-modelos que intentan predecir la curvatura de una molécula de DNA con la mayor precisión posible. Algunos de los más importantes son:

- Modelo de De Santis (3). Los valores de las variables geométricas del DNA se obtienen de cálculos de minimización de la energía conformacional de la molécula de DNA, verificados mediante el análisis de cristales de oligonucleótidos, y la correcta predicción de la movilidad de fragmentos de DNA curvos en geles de agarosa.
- Calladine et al (4) generaron un modelo de predicción de regiones curvas basados en datos de difracción de rayos X de algunos oligos, especialmente conformados por secuencias de poliamina flanqueadas por otros nucleótidos, en este modelo el dinucleótido *a-a* es considerado recto (con valores angulares de 0), el dinucleótido *a-t* curvo, con un ángulo de roll de 6 grados y el resto de los dinucleótidos son considerados "ligeramente doblados" con un ángulo de roll constante de 3 grados.
- Modelo de Bolshoi (5). Este modelo es generado a partir del análisis de 54 oligonucleótidos sintéticos, da una contribución importante a la variable tilt, y permite un gran intervalo de variación en la variable twist, que otros modelos consideran constante.
- Posicionamiento de nucleosomas (6). Las contribuciones angulares de este modelo son derivadas exclusivamente de observaciones experimentales sobre la preferencia de la ubicación de trímeros en círculos pequeños de DNA y regiones enrolladas en nucleosomas. A

partir de estas secuencias fueron obtenidas las frecuencias de ocurrencia de cada trímero en el lado cóncavo (el surco mayor) del DNA. Valores de roll fueron asignados en base a estas frecuencias. Este modelo ha demostrado ser el más exacto en cuanto a su capacidad de predicción de la estructura de oligómeros y su comportamiento en geles de agarosa, y es el modelo utilizado en el presente estudio.

El papel de la curvatura del DNA en la biología.

La relevancia biológica de la curvatura del DNA ha sido estudiada por más de 20 años, y se ha relacionado con una gran variedad de procesos biológicos, tales como la replicación del DNA, recombinación, transposición, empaquetamiento de la cromatina y regulación transcripcional (7). Estos estudios han sido realizados únicamente en *loci* específicos y regiones discretas de DNA de algunos organismos modelo.

El desarrollo en la tecnología de secuenciación de DNA ha generado un crecimiento exponencial en el tamaño de las bases de datos de secuencia, y se ha determinado la secuencia completa de más de 100 genomas arqueobacterianos, y al rededor de 10 genomas eucariontes. Este caudal de información ha abierto la oportunidad de realizar estudios que contemplan al genoma como unidad fundamental, y permiten comparaciones globales entre diferentes genomas.

Determinantes globales de curvatura.

El perfil de curvatura de un genoma completo fue descrito por primera vez en 1997 (8) y se demostró que cada genoma posee un perfil de curvatura característico. Nuestro trabajo extendió este análisis caracterizando los determinantes biológicos de la curvatura del DNA de genomas totales, mediante la manipulación *in silico* de la secuencia nucleotídica.

Curvatura del DNA y regulación transcripcional.

El papel de la curvatura del DNA como elemento regulador de la transcripción ha sido confirmado para numerosos genes en donde se ha demostrado que las regiones curvas favorecen la interacción entre el DNA y la RNA polimerasa, así como la formación del complejo abierto (12, 13). Existen también reportes que indican que algunas proteínas reguladoras, tales como IHF, H-NS, FNR, FIS y HU reconocen sitios curvos en el DNA (14, 15). Recientemente se han iniciado análisis globales de la curvatura de las regiones promotoras en genomas completos, que han identificado tendencias a valores altos de curvatura en genomas bacterianos y de fagos, comparados con genomas eucariontes, virales y mitocondriales (16). También se han caracterizado regiones de regulación de alta curvatura promedio en genomas de bacterias mesófilas (17) y señales de DNA curvo en algunos promotores de micobacterias (18).

Desarrollo.

La metodología seguida en este proyecto, así como los principales resultados, fueron publicados en 2 artículos en revistas de arbitraje internacional (Microbial and Comparative Genomics), los cuales se presentan aquí. La última parte del desarrollo de este proyecto, que involucra el análisis de regiones de regulación, fue reportada en un manuscrito, sometido para su publicación (Nucleic Acids Research). El manuscrito en su totalidad compone la última parte del desarrollo.

MICROBIAL & COMPARATIVE GENOMICS
Volume 3, Number 4, 1998
Mary Ann Liebert, Inc.

Relationship between Codon Usage and Sequence-Dependent Curvature of Genomes

RUY JÁUREGUI,[1] FEDERICO O'REILLY,[2] FRANCISCO BOLIVAR,[3] and ENRIQUE MERINO[3]

ABSTRACT

Static DNA curvature distributions of full-sequenced genomes and large DNA contigs from different organisms were calculated. Very distinctive differences among histogram profiles coming from archaebacteria, eubacteria, and eukaryotes were observed. Eubacterial profiles were, on average, more curved than were archaeal and eukaryotic profiles. A comparative analysis between real and randomized DNA sequences revealed that eubacterial genomes presented, overall, higher curvature values than random sequences. An opposite portrait was exhibited by archaeal and eukaryotic genomes. They displayed a lower frequency of curved regions than their corresponding randomized sequences. The contributions of coding and intergenic regions to the curvature profile were also analyzed. Intergenic regions, on average, were found to be more curved than the overall genomic sequences, especially in prokaryotic organisms. Nevertheless, because of their small size with respect to coding regions, the contribution of intergenic sequences to the overall curvature profile tended to be minor. A clear relationship between codon usage and DNA curvature was demonstrated, and a proposal of the possible coevolution of both systems is discussed. Finally, we present a procedure to quantify the deviation of a curvature profile from randomness through a formal statistical analysis.

INTRODUCTION

The relevance of DNA curvature as a regulatory component in transcription, replication, recombination, and chromatin structure has been proposed for more than 15 years (reviewed by Hageman, 1990; Harrington, 1992; Travers, 1990; Kathleen, 1992; Pérez-Martín et al., 1994). The increasing evidence in favor of DNA sequence-dependent conformations has encouraged the study of the bending propensity parameters of dinucleotides and trinucleotides. Bending contribution matrices have been deduced from analysis of different sequences by electrophoresis gel retardation (Calladine et al., 1988; Bolshoy et al., 1991), x-ray diffraction (Koo and Crothers, 1988; Nelson et al., 1987), NMR analysis (Clore and Gronenborn, 1985; Sarma et al., 1988), DNAse protection experiments (Satchwell et al., 1986), and, theoretically, energy calculations (Cacchione et al., 1989; De Santis et al., 1986). These data, in conjunction with computer algo-

[1]Laboratorio de Biología Computacional. Centro de Investigación sobre Fijación de Nitrógeno, [2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, and [3]Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México.

rithms, have been used mainly to study the curvature of short DNA fragments at specific and discrete loci. The development of new and better nucleotide sequencing techniques offers the possibility of extending these studies from restricted and very short DNA regions into a novel field that could provide different insights into the biologic role of DNA curvature by analysis and comparison of complete genomes or at least of very large DNA contigs. Pioneer research in this new field is the work of Gabrielian et al. (1997), who studied the distribution of sequence-dependent curvature in prokaryote and eukaryote genomes, showing that prokaryotic genomes present a higher content of curved DNA than eukaryote genomes.

Here, we extend this insight by analysis of recently sequenced genomes, including those from the archaeal domain, taking into account new variables that may affect DNA curvature, such as the presence of coding or intergenic sequences and the codon usage bias. The effect of this bias was analyzed by generation of sequences in which *Escherichia coli*, *Methanococcus jannaschii*, and *Arabidopsis thaliana* DNA coding regions were set up based on the codon usage of different organisms. Local curvature values at every nucleotide were grouped into histogram profiles. Significant changes in the DNA curvature profiles were found, depending on the codon usage applied. Our results also suggest that the DNA curvature of genomic sequences is a nonfortuitous phenomenon. Comparison of DNA curvature profiles showed important differences between real and shuffled sequences. Eubacterial chromosomes showed a higher content of curved regions than random sequences, and archaeal and eukaryotic genomes exhibited a lower frequency of curved regions. A formal method to evaluate the statistical significance of the sequence-dependent curvature of chromosomal DNA is presented, and the results obtained and their implications are discussed.

MATERIALS AND METHODS

The BEND program of Goodsell and Dickerson (1994) was translated into Perl language and adapted to make the analysis of whole genomes or large contig sequences. The BEND program is based on the nearest-neighbor model, which assumes that curvature of DNA is the result of successive accumulation of rotational and spatial displacement between base pairs. The program reads the DNA sequence, evaluates the normal vector of each base pair, and averages these values over a 10 bp interval. Curvature is then calculated as the angle between averaged normal vectors 31 bp away. To have a standard unit of curvature, we express the curvature value as the deviation angle per one helical turn (10.5 nt). Among different bending contribution matrices published, we chose the trimer matrix of the nucleosome position model of Satchwell et al. (1986) because it has been used to accurately predict the curvature of well-characterized curved DNA sequences.

Randomization of the sequences was made by the random repositioning of every base or set of bases along the entire nucleotide sequence. When the randomization window was larger than 1 nt, we divided the whole sequence into N windows (where N = sequence length/window size). Then, every window was randomly repositioned along the sequence. Programs were executed in Sun Ultra and Silicon Graphics PowerChallenge computers under UNIX platform. The program sources and examples are available at our website, http://www.ibt.unam.mx/cgi-bin/server/PRG.base?clase:iap.

The sequence files of the genomes of *A. thaliana*, *Archeoglobus fulgidus*, *Borrelia burgdorferi*, *Bacillus subtilis*, *Chlamydia trachomatis*, *E. coli*, *Haemophilus influenzae Helicobacter pylori*, *Methanolabacterium thermoautotrophicum*, *Mycoplasma genitalium*, *M. jannaschii*, *Mycoplasma pneumoniae*, *Saccharomyces cerevisiae*, and *Synechocystis* sp. were retrieved from The Institute for Genomic Research (TIGR) worldwide webserver (http://www.tigr.org/tdb/mdb/mdb.html). *Plasmodium falciparum*, *Caenorhabiditis elegans*, and *Homo sapiens* were retrieved from The Sanger Center (http://www.sanger.ac.uk).

RESULTS

Curvature profiles of real and randomized sequences

To analyze the curvature of genomic DNA sequences, we have implemented a program based on the BEND algorithm developed by Goodsell and Dickerson (1994) that calculates the intrinsic local curvature of B-DNA tracts. We used a set of parameters that have been shown to predict accurately the curvature of well-characterized curved DNA sequences (Goodsell and Dickerson, 1994). Evaluation of local DNA curvature was determined at each nucleotide for the following full-sequenced genomes: *A. fulgidus*, *B. burgdorferi*, *E. coli*, *B. subtilis*, *H. influenzae*, *H. pylori*, *M. thermoautotrophicum*, *M. genitalium*, *M. jannaschii*, *M. pneumoniae*, *S. cerevisiae*, and *Synechocystis* sp. Large DNA contigs from *C. elegans*, *C. trachomatis*, *A. thaliana*, *Drosophila melanogaster*, *P. falciparum*, and *H. sapiens* were also included in our analysis.

Length and GC content of the sequences analyzed, as well as the mean of the curvature values, are listed in Table 1. Local curvature values at every nucleotide were grouped into histogram profiles with a bin width of 0.33 degrees per helical turn. Frequencies are normalized and expressed as a percentage of the size of the analyzed DNA. With few special exceptions, eubacterial genomes had a higher content of curved DNA sequences than archaeal or eukaryotic genomes. This result is in good agreement with the data obtained by Gabrielian et al. (1997). *C. trachomatis* was the most important exception to this rule (Fig. 1). Its DNA showed the lowest curvature value among the eubacterial sequences studied and displayed a curvature profile less curved than that of *C. elegans*, *D. melanogaster*, and *S. cerevisiae* DNA. Most of the eubacterial genomes analyzed showed a remarkable similarity in their curvature profiles (represented as a single curvature profile in Fig. 1, thick line), with the distinctive exceptions of the *H. influenzae* and *H. pylori* genomes, which presented the highest curvature values, and the *C. trachomatis* DNA just described. A wider spread was observed between the archaeal or eukaryotic genomes. *M. jannaschii* and *M. thermoautotrophicum* were the organisms from the Archaea domain that displayed the highest and lowest DNA curvature distributions, respectively, whereas in the Eukarya domain, *C. elegans* and *H. sapiens* displayed the highest and lowest DNA curvature distributions, respectively (Fig. 1).

TABLE 1. COMPILATION OF CURVATURE AND STATISTICAL VALUES OF GENOMES AND LARGE DNA CONTIGS

Organism	Domain	Size (Kbp)	GC%	Curvature of real sequences [a]	Curvature of random sequences [b]	DSDZ [c]	$DSDZ_{500}$ mean value [d]
Bacillus subtilis	Eubacteria	4,215	43.52	3.94	3.61	66.42	23.64
Borrelia burgdorferi	Eubacteria	812	29.85	3.96	3.78	19.67	11.59
Chlamydia trachomatis	Eubacteria	1,042	41.31	3.74	3.67	14.61	7.15
Escherichia coli	Eubacteria	4,639	50.79	3.92	3.46	101.11	33.74
Haemophilus influenzae	Eubacteria	1,830	38.15	4.37	3.70	88.62	46.09
Helicobacter pylori	Eubacteria	1,668	38.87	4.80	3.70	132.81	77.22
Mycoplasma genitalium	Eubacteria	581	31.69	4.00	3.77	17.70	16.58
Mycoplasma pneumoniae	Eubacteria	816	40.01	3.98	3.77	25.48	16.13
Synechocystis sp.	Eubacteria	3,574	47.72	4.05	3.52	97.09	37.04
Archaeoglobus fulgidus	Archaea	1,974	48.12	3.70	3.50	27.45	13.34
Methanococcus jannaschii	Archaea	1,665	31.43	3.93	3.78	20.76	11.46
Methanobacterium thermoautotrophicum	Archaea	1,594	49.08	3.23	3.49	37.97	22.13
Arabidopsis thaliana	Eucaryote	1,036	36.26	3.60	3.89	13.93	10.54
Caenorhabditis elegans	Eucaryote	31,888	35.39	4.06	3.74	221.82	22.13
Drosophila melanogaster	Eucaryote	7,978	42.46	3.88	3.63	70.82	18.06
Homo sapiens	Eucaryote	2,583	44.70	3.33	3.60	43.38	19.55
Plasmodium falciparum	Eucaryote	11,137	22.82	3.42	3.73	115.63	25.53
Saccharomyces cerevisiae	Eucaryote	12,069	38.30	3.78	3.70	27.98	6.94

[a] Mean of the curvature value of the nucleotides in the original sequence. The curvature value at every base is expressed as the deviation angle per 10.5 nt.

[b] Mean of the curvature value of the nucleotides in the randomized sequences.

[c] Distance, in standard deviations, between zero (DSDZ) and the mean of the area values. See text for complete description.

[d] The sequence is divided in blocks of 500 kbp, and the mean of their DSDZ values ($DSDZ_{500}$) is expressed.

To support the nonfortuitous nature of the curvature profiles, we generated new nucleotide sequences by random shuffling of the original genomic sequences. Randomized sequences have the same size and GC content as the original genomes but differ in the relative positions of their nucleotides. Curvature profiles of all randomized sequences were almost identical regardless of their eubacterial, archaeal, or eukaryotic origin. The curvature profiles of these randomized genomes can be retrieved from our webserver (http://www.ibt.unam.mx/cgi-bin/server/PRG.base?clase:iap). We found that most of the eubacterial sequences have a higher content of curved regions than their randomized sequences. *H. pylori* curvature profiles are illustrated in Figure 2A as an example. Gabrielian et al. (1997) found similar results in a study of the *E. coli* genome. As this increased content of curved regions in real sequence was a common outcome, an important finding was the existence of some archaeal and eukaryotic genomes that displayed a lower content of curved DNA than those exhibited by randomized sequences. Figure 2B and 2C show the profiles of the archaebacterium *M. thermoautotrophicum* and the eukaryote *H. sapiens*, respectively. Table 1 lists the mean of the curvature values of each genome analyzed, as well as the values obtained from their corresponding randomized sequences.

In an attempt to search for the minimal contribution unit of curvature, we expanded the randomizing window from 1 nt to 2, 3, 5, 10, 20, and so on up to 100 nt (see Materials and Methods). We looked for the minimum window size that generates no differences between real and randomized genomes. We found that shuffling sequences with a window size of 100 base creates almost the same profile as the original genome. Window sizes between 1 and 100 bp produced randomized profiles that gradually approached the profile of the real genome in a fairly geometric fashion. The DNA curvature profiles of this study can be retrieved from our webserver (http://www.ibt.unam.mx/cgi-bin/server/PRG.base?clase:iap).

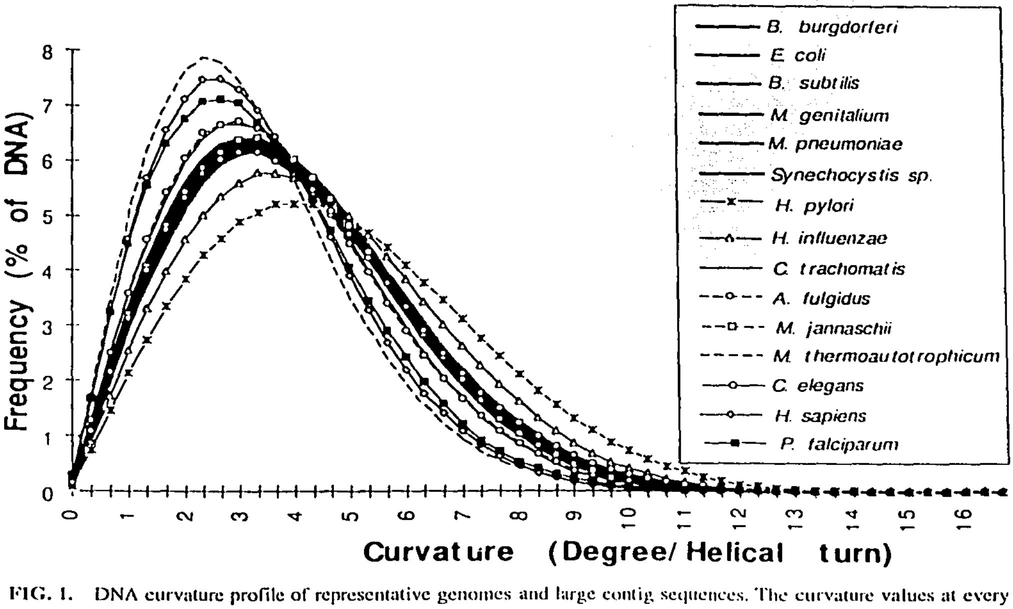


FIG. 1. DNA curvature profile of representative genomes and large contig sequences. The curvature values at every nucleotide of a DNA sequence are grouped into bins of a width that corresponds to 0.33 degreees per helical turn. Frequencies are normalized and expressed as a percentage of the DNA analyzed. The thick line includes the profiles of *B. burgdorferi, E. coli, B. subtilis, M. genitalium, M. jannaschii, M. pneumoniae,* and *Synechocystis* sp. For clarity, only representative examples of the sequences analyzed are included. Full color graphics, including the results of all the organisms analyzed, are accessible from http://www.http://www.ibt.unam.mx/cgi-bin/server/PRG.base?clase:iap.

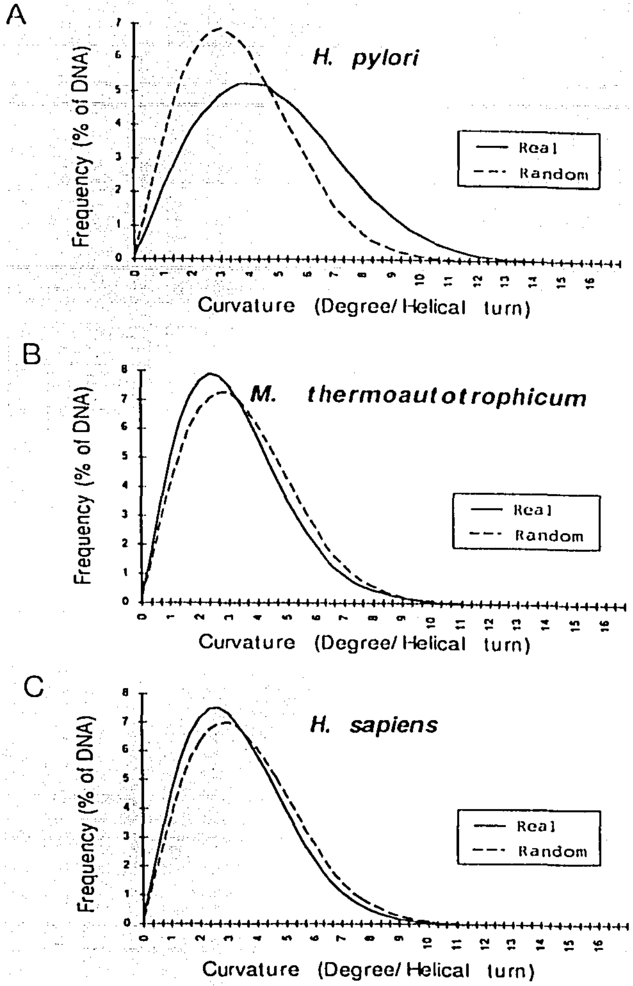


FIG. 2. Comparison of DNA curvature profiles from real and randomized sequences. (A) Eubacterial, (B) archaeal, and (C) eukaryotic DNA curvature profiles, respectively, where we found the major differences in the curvature values between real and randomized sequences.

Contribution of coding and intergenic sequences to the overall curvature profile

When coding information was available, we performed a comparative analysis of the curvature profile of coding and intergenic regions (Fig. 3). Although curvature profiles exhibited similar overall shapes, curvature profiles of intergenic regions showed a deviation at their right end tails. Such a deviation corresponds to an increase in the high curvature value frequencies (Fig. 3A, inset). In all cases, the curvature profiles of intergenic sequences showed higher curvature values than the profiles of coding regions or the entire genomes. The extent of these differences varied importantly depending on the organism analyzed. Nevertheless, in all the organisms studied, the curvature profiles of the coding regions were slightly smaller than those of entire genomes. Curvature profiles of representative examples of eubacterial (*E. coli*), archaeal (*A. Fulgidus*), and eukaryotic (*S. cerevisiae*) organisms are shown in Figure 3.

Effect of codon usage bias on DNA curvature

A possible bias on DNA curvature due to codon usage was also examined. The *E. coli, M. jannaschii*, and *A. thaliana* genomes were chosen as examples of eubacterial, archeal, and eukaryote genomes. The codon usage tables of these organisms, as well as those *H.pylori* and *M. thermoautotrophicum*, were calculated based on their annotated genomic sequences. *H. pylori* and *M. thermoautotrophicum* represent the

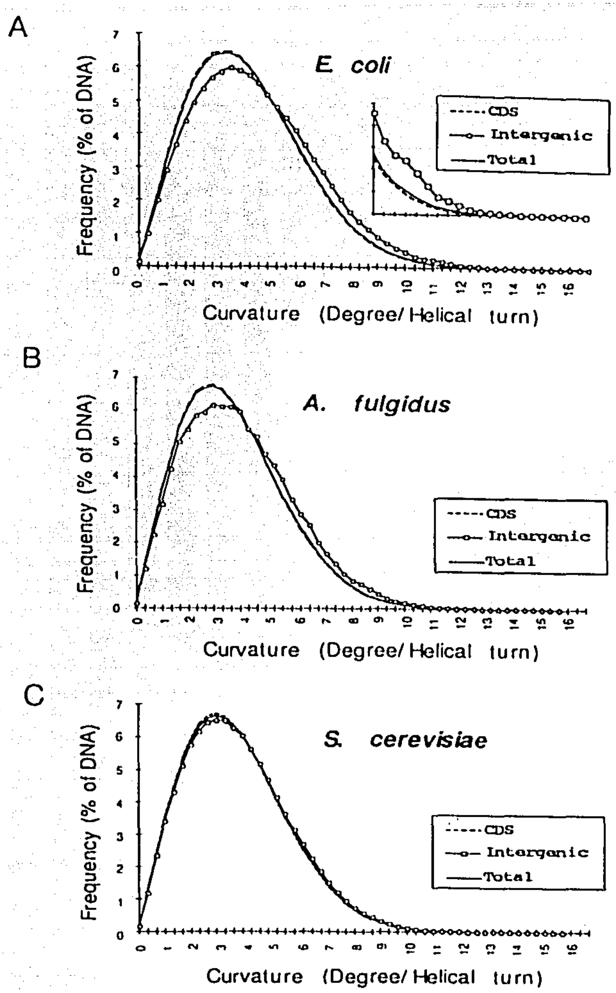


FIG. 3. DNA curvature profiles of coding, intergenic, and overall DNA genomic sequences. (A) Eubacterial, (B) archaeal, and (C) eukaryotic curvature profiles.

organisms with the highest and lowest content, respectively, of curved DNA, as previously described (Fig. 1 and Table 1). We generated three different *E. coli* databases in which the triplets of every coding region were replaced with a synonym triplet according to the codon usage bias of either *E. coli, H. pylori*, or *M. thermoautotrophicum*. Significant differences in the curvature profiles were displayed. The profiles with the highest and lowest curvature values corresponded to the *E. coli* artificial genomes generated with *H. pylori* and *M. thermoautotrophicum* codon usage tables, respectively (Fig. 4A). Similar results were ob-

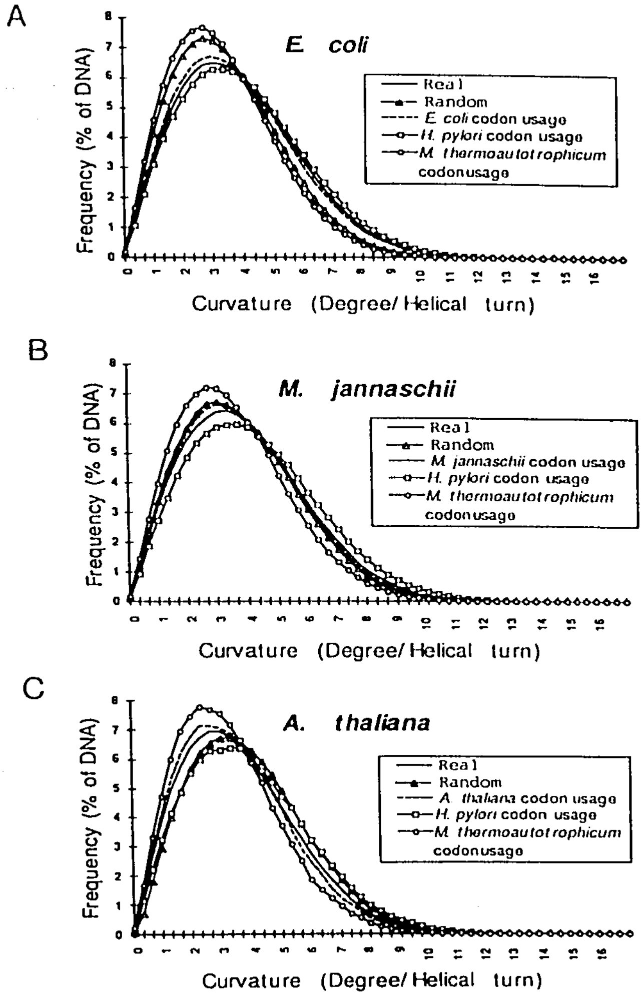


FIG. 4. Effect of codon usage bias on DNA curvature. Artificial *E. coli, M. jannaschii*, and *A. thaliana* DNA sequences were created based on the codon usage of organisms with either the highest (*H. pylori*) or lowest (*M. thermoautotrophicum*) content of curved DNA, and their curvature profiles were calculated. As internal controls, the artificial *E. coli, M. jannaschii*, and *A. thaliana* DNA sequences generated on the basis of their own codon usage bias were analyzed. The curvature profiles of the original and randomized genomes are also presented.

tained with the genome sequences of *M. jannaschii* and *A. thaliana* (Fig. 4B and 4C, respectively). Our internal controls were the *E. coli*, *M. jannaschii*, and *A. thaliana* databases generated with their own codon usage bias. The curvature profiles of these new genomes were much closer to the profiles obtained with the real *E. coli*, *M. jannaschii*, and *A. thaliana* sequences than those obtained with their correspondent randomized genomes. These results indicate that there is, in fact, a relationship between codon usge and DNA curvature. Nevertheless, other factors also might be related. Similar results were obtained with the codon usage tables of *H. influenzae*, *M. genitalium*, *M pneumoniae*, and *S. cerevisiae* and are available at our WWW site (http://www.ibt.unam.mx/cgi-bin/server/PRG.base?clase:iap).

Statistical significance of DNA curvature

We developed a procedure to estimate the statistical significance of the curvature profile of entire genomes or large contig DNA sequences. In this procedure, we generated, by shuffling, a large number of randomized versions of the original genome and calculated the curvature profiles of each randomized sequence. Differences between the curvature histogram of each randomized version and the original sequence were evaluated bin by bin in order to obtain the area values between the real profile and each of the randomized profiles. From analysis of 1000 randomized sequences, values closely following a normal distribution were obtained, and their mean and SD were calculated. Based on the fact that the area between two identical profiles is zero, the probability of a random sequence having the same curvature profile as the original can be computed from the distance, in standard deviations (SD), between zero (DSDZ) and the mean of the area values. A DSDZ >4 is highly unlikely to occur in a normal distribution. All genomes proved to have very significant DSDZ values, regardless if they were deviated toward a more curved (eubacteria, in general) or less curved (most of the archaebacteria and eukaryotes studied) profile, than those obtained by their corresponding random sequences (Fig. 2 and Table 1). A significant influence toward high DSDZ values were observed as a result of the size of the sequences analyzed. To make fair comparisons, we chose 500 kbp as our standard length of analysis. This size approximately corresponds to the smallest size of the genomes studied (*M. genitalium*, 580,820 bp). We called the DSDZ value obtained from the analysis of a 500 kbp window $DSDZ_{500}$. In general, all of the 500 kbp fragments coming from one genome or DNA contig presented almost identical curvature profiles and $DSDZ^{500}$ values. It is important to note that eubacterial genomes showed the most significant $DSDZ_{500}$ values. In our study, *H. pylori*, *M. thermoautotrophicum*, and *P. falciparum* were the eubacteria, archaebacteria, and eukaryote with the highest $DSDZ_{500}$ values (77, 22, and 25 $DSDZ_{500}$, respectively) (Table 1).

DISCUSSION

What kind of new biologic information can be obtained from systematic analysis of the DNA sequences reported in the rapidly growing databases? Can old questions, restricted to discrete DNA regions, be rephrased to large DNA contigs or even entire genomes? Sequence-dependent DNA curvature is one example of a fundamental issue that can be reanalyzed in this era of massive nucleotide sequence acquisition. In this study, we showed that the intrinsic curvature of chromosomal DNA presents distinctive profiles that may be related to their eubacterial, archaeal, or eukaryotic origin. Important common properties are shared by Archaea and Eukarya, such as their tendency to have a low content of curved DNA sequences and the great variation between their DNA curvature profiles. On the other hand, eubacterial genomes tend to be more curved and show an exceptional similarity in their DNA curvature profiles. These results are in good agreement with the work of Gabrielian et al. (1997), which showed that some prokaryotic genomes (*H. influenzae*, *M. jannaschii*, and *M. genitalium*) appear to have a higher frequency of curved DNA than eukaryotic genomic DNA sequences (*H. sapiens* and *S. cerevisiae*). Although both studies reached the same conclusion, they differ in the extent of DNA curvature observed because of the characteristics of the geometric matrices used in each analysis. We applied the trimer matrix of the nucleosome position model (Satchwell et al., 1986), whereas Gabrielian et al. (1997) used the geometric parameters of the NMR model (Ulyanov and James, 1995).

From our results, it is inferred that the overall curvature of genomic DNA is a nonfortuitous phenomenon and possesses special characteristics that are not found in random sequences. Elucidation of the driving forces that determine the DNA curvature is one of the main goals of our study. In this regard, we consider the constraints encountered by DNA to be organized and packed in the chromosome as an important element to study. Evidence supporting the relationship between DNA curvature and chromosomal structure is based on the periodicities of some dinucleotides present in the DNA nucleosome core at every 10.5 bp (the size of a DNA helical turn) (Satchwell et al., 1986; Trifonov and Sussman, 1980). It is important to note that the archaeal chromosomal DNA is histone associated and organized into nucleosome-like structures. Such characteristics suggest an organization similar to that found in eukaryotes (Takayanagi et al., 1992). In that respect, it is possible that the nucleosome organizations in Archaea and Eukarya are closely related to their low content of curved sequences in their genomes, whereas Eubacteria have evolved a less structured but more curved genome.

To provide additional support to this hypothesis, we have analyzed the curvature profile of the eubacterium *C. trachomatis*, an obligate intracytoplasmic parasite of eukaryotic cells that has a condensed nucleoid and a eukaryotic H1-like protein (Costerton et al., 1976). We observed that the curvature profile of *C. trachomatis* DNA is notably different from the rest of the eubacterial curvature profiles and resembles archaeal and eukaryotic profiles, supporting the idea of a relationship between chromosomal structure and static DNA curvature. Interestingly, Miramontes et al. (1995) were also able to distinguish prokaryotic from eukaryotic DNA base on local stacking and structural properties of DNA as a function of the binary distributions of GC/AT and purine/pyrimidine base pairing. Supported on these parameters, these authors also reached the conclusion that the requirements of chromosomal DNA to be packed impose specific constraints in DNA sequences that could reveal different evolutionary DNA histories.

The influence of curved DNA regions on gene transcription regulation in prokaryotic and eukaryotic organisms has been well established (Hagerman, 1990; Párez-Martín et al., 1994; Schatz and Langowski, 1997). We expected to find a higher content of curved DNA sequences in the intergenic regions because of their regulatory nature and corresponding influence on the overall DNA curvature. Therefore, we analyzed the contributions of coding and intergenic regions in the curvature of the genomic DNA. Actually, the anticipated enhanced frequency of curved DNA sequences in the intergenic regions was found, but this enrichment had only a minor impact on the overall chromosomal DNA curvature, as their size is comparatively small with respect to the size of the coding sequences.

An interesting observation is the remarkable similarity between the curvature profiles of the coding and the overall genomic sequences, implying that the DNA constraints influencing DNA curvature are notably present in coding regions. In agreement with this conclusion and considering that codon usage is an important feature of the coding regions, we envisioned a possible relationship between the codon usage bias and the curvature of genomic DNA sequences. To test this premise, we generated hypothetical *E. coli*, *M. jannaschii*, and *A. thaliana* genomes that presented the codon usage pattern of other organisms. Interestingly, we found that artificial genomes with the codon usage of organisms with either a high (*H. pylori*) or a low (*M. thermoautotrophicum*) content of curved DNA exhibited a correspondingly high or low frequency of curved DNA sequences. Similar results were obtained with the codon usage of different organisms. These data clearly demonstrate a significant interdependence between codon usage and DNA curvature. It has been reported that preferential codon usage could be a consequence of the relative abundance of isoaccepting tRNA (Bulmer, 1987; Ikemura, 1985), translational selection (Xia, 1996), or mutational pressure in DNA (Sharp et al., 1993). Here, we present evidence to consider that DNA curvature is another element related to codon usage. Nevertheless, from our results, it is not possible to determine whether organisms adopted a particular codon usage during evolution in order to obtain a specific DNA curvature or, conversely, a selection pressure for a specific codon usage had an indirect effect on DNA curvature. Because both the preference of codon usage and the curvature in chromosomal DNA seem to be important requirements for the biology of every organism, a plausible scenario could have included a coevolution of both systems. However, validation of this novel proposal will require additional analysis.

We developed a simple procedure to evaluate how far the genomic DNA curvature is from random. For that purpose, we generated 1000 randomized versions of the original genomic sequence, and for each of them, we evaluated the area between the curvature profiles of the real and randomized sequences. We obtained data clearly following a normal distribution of these 1000 area values and calculated the distance, in

SD, between zero and the mean of the area values. For all the organisms analyzed, this distance was shown to be clearly different from zero.

We found our statistical procedure to be a useful tool to corroborate our graphic observations and to assess their statistical significance as to the nonrandom nature of the DNA curvature. Although the procedure described was specifically designed to analyze the nonrandom nature of DNA curvature, it can also be used to study some other global variables of genomic DNA, such as GC content, codon usage, and dinucleotide distribution. The high significant values obtained in our statisitical analysis evidence the importance of DNA curvature not only for discrete and restricted regions but also for overall DNA genomic sequences.

ACKNOWLEDGMENTS

We thankfully acknowledge the vital help of Mara Tellez and Jorge Olguín in their advice on the statistical algorithm. We are indebted to Rosana Sánchez for continued interest in this project, her thoughtful advice, and critical reading of the manuscript. We also thank Edmundo Calva, Julio Collado, Alejandro García, José Luis Puente, Lorenzo Segovia, and Roberto Stock for critical reading of the manuscript and Ricardo Ciria, Abel Linares, Arturo Ocadis, and Alma Martínez for computer support.

REFERENCES

BOLSHOY, A., McNAMARA, P., HARRINGTON, R.E., and TRIFONOV, E.N. (1991). Curved DNA without A-A: Experimental estimation of all 16 DNA wedge angles. Proc Natl Acad Sci USA **88**, 2312–2316.

BULMER, M. (1987). Coevolution of codon usage and transfer RNA abundance. Nature **325**, 728–730.

CACCHIONE, S., De SANTIS, P., FOTI, D., PALLESCHI, A., and SAVINO, M. (1989). Periodical polydeoxynucleotides and DNA curvature. Biochemistry **28**, 8706–8713.

CALLADINE, C.R., DREW, H.R., and McCALL, M.J. (1988). The intrinsic curvature of DNA in solution. J Mol Biol **201**, 127–137.

CLORE, G.M., and GRONENBORN, A.M. (1985). Probing the three-dimensional structures of DNA and RNA oligonucleotides in solution by nuclear Overhauser enhancement measurements. FEBS Lett **179**, 187–198.

COSTERTON, J.W., POFFENROTH, L., WILT, J.C., and KORDOVA, N. (1976). Ultrastructural studies of the nucleoids of the pleomorphic forms of *Chlamydia psittaci* 6BC: A comparison with bacteria. Can J Microbiol **22**, 16–28.

De SANTIS, P., MOROSETTI, S., PALLESCHI, A., and SAVINO, M. (1986). In *Structure and Dynamics of Nucleic Acids, Proteins and Membranes*. E. Clementi and S. Chim, eds. (Plenum Publishing Corp., New York), 31–49.

GABRIELIAN, A., VLAHOVICEK, K., and PONGOR, S. (1997). Distribution of sequence-dependent curvature in genomic DNA sequences. FEBS Lett **406**, 69–74.

GOODSELL, D.S., and DICKERSON, R.E. (1994). Bending and curvature calculations in B-DNA. Nucleic Acids Res **22**, 5497–5503.

HAGERMAN, P.J. (1990). Sequence-directed curvature of DNA. Annu Rev Biochem **59**, 755–781.

HARRINGTON, R.E. (1992). DNA curving and bending in protein-DNA recognition. Mol Microbiol **6**, 2549–2555.

IKEMURA, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol **2**, 13–34.

KATHLEEN, S.M. (1992). DNA looping. Microbiol Rev **56**, 123–136.

KOO, H.-S., and CROTHERS, D.M. (1988). Calibration of DNA curvature and a unified description of sequence-directed bending. Proc Natl Acad Sci USA **85**, 1763–1767.

MIRAMONTES, P., MEDRANO, L., CERPA, C., CEDERGREN, R., FERBEYRE, G., and COCHO, G. (1995). Structural and thermodynamic properties of DNA uncovered different evolutionary histories. J Mol Evol **40**, 698–704.

NELSON, H.C., FINCH, J.T., LUISI, B.F., and KLUG, A. (1987). The structure of an oligo(dA) oligo(dT) tract and its biological implications. Nature (Lond) **330**, 331–226.

PÉREZ-MARTIN, J., ROJO, F., and De LORENZO, V. (1994). Promoters responsive to DNA bending: A common theme in prokaryotic gene expression. Microbiol Rev **58**, 268–290.

SARMA, M.H., GUPTA, G., and SARMA, R.H. (1988). Structure of a bent DNA: Two-dimensional NMR studies on d(GAAAATTTTC). Biochemistry **27**, 3423–3432.

SATCHWELL, S.C., DREW, H.R., and TRAVERS, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. J Mol Biol **191**, 659–675.

SCHATZ, T., and LANGOWSKI, J. (1997). Curvature and sequence analysis of eukaryotic promoters. J Biomol Struct Dyn **15**, 265–275.

SHARP, P.M., STENICO, M., PEDEN, J.F., and LLOYD, A.T. (1993). Codon usage: Mutational bias, translational selection, or both? Biochem Soc Trans **21**, 835–841.

TAKAYANAGI, S., MORIMURA, S., KUSAOKE, H., YOKOYAMA, Y., KANO, K., and SHIODA, M. (1992). Chromosomal structure of the halophilic archaebacterium *Halobacterium salinarium*. J Bacteriol **174**, 7207–7216.

TRAVERS, A.A. (1990). Why bend DNA? Cell **60**, 177–180.

TRIFONOV, E.N., and SUSSMAN, J.L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci USA **77**, 3816–3820.

ULYANOV, N.B., and JAMES, T.L. (1995). Statistical analysis of DNA duplex structural features. Methods Enzymol **261**, 90–120.

XIA, X. (1996). Maximizing transcription efficiency causes codon usage bias. Genetics **144**, 1309–1320.

Address reprint requests to:
Dr. Enrique Merino
Departmento de Microbiología Molecular
Instituto de Biotecnología
UNAM
Apdo.Postal 510-3
Cuernavaca, Morelos Cpb2271
México
E-mail: merino@ibt.unam.mx

MICROBIAL & COMPARATIVE GENOMICS
Volume 5, Number 1, 2000
Mary Ann Liebert, Inc.

Relationship between Whole Proteome Aminoacid Composition and Static DNA Curvature

RUY JÁUREGUI, FRANCISCO BOLIVAR, and ENRIQUE MERINO

ABSTRACT

To study possible relationships between an organism's *genomic DNA curvature* and the *aminoacid composition* of its *proteome*, every peptidic sequence from fully determined genomes was retrotranslated using the *E. coli* codon preferences, and the curvature profiles of the resulting DNA sequences were calculated and compared. A clear interdependence between these two variables was observed, as each retrotranslated proteome presented a distinctive, statistically significant DNA curvature profile biased toward its natural DNA curvature profile. In addition, by comparing the profiles arising from real and randomly permuted proteomes, we also found a position-dependent contribution of the peptidic sequence to DNA curvature. The implications of these results support the idea of a possible selection toward a specific global curvature of genomes.

INTRODUCTION

THE EARLIEST STRUCTURAL CHARACTERIZATIONS OF THE DNA MOLECULE described it as a double helix following an essentially linear trajectory, supposing that the base pairs were parallel to each other (Watson and Crick, 1953). Further studies revealed the existence of electromagnetic interactions that modify the position of a base pair with respect to its neighbors in a sequence-dependent way. Several topological variables have been defined to describe the deviations in the base pair stacking due to these interactions, such as twist, slide, tilt and roll variables (for a review see Dickmann, 1989). The overall effect of these interactions is to deviate the trajectory of the DNA molecule from an ideal straight line, and depending on the sequence, in some cases produce a curved trajectory.

Biological roles for static DNA curvature have been well documented for a number of discrete loci that include a broad spectrum of biological functions, such as transcription regulation, replication, recombination and chromatin structure (Pérez-Martin et al., 1994; Hagerman, 1990; Harrington, 1992; Travers, 1990). It was not until the advent of more powerful DNA sequencing technologies, and the following exponential growth on sequence databases, that it became possible to make whole-genome characterizations of DNA curvature. The pioneering work of Gabrielian et al. (1997) demonstrated that every organism possesses a characteristic DNA curvature profile. Further work of our group demonstrated a statistically significant relationship between DNA curvature and codon usage, by the analysis of DNA curvature profiles coming from the retrotranslation of a single proteome using codon preferences from different organisms (Jáuregui et al., 1998).

Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México.

On the other hand, biases on aminoacid composition of proteomes have been related to GC content (Cole et al., 1998), dinucleotide frequencies (Karlin et al., 1997), and the position of the genes at the leading or lagging strand (Mrázek and Karlin, 1998).

Here we describe the relationship between aminoacid composition of proteomes and genomic DNA curvature, and discuss its possible implications.

MATERIALS AND METHODS

Artificial DNA sequences were generated by retrotranslating the complete proteomes obtained from the annotations of the GenBank files of the following organisms: *Aquifex aeolicus, Aeropyrum pernix, Archaeoglobus fulgidus, Borrelia burgdorferi, Bacillus subtilis, Chlamydia pneumoniae, Chlamydia trachomatis, Escherichia coli, Caenorhabditis elegans (chromosome 5), Haemophilus influenzae, Helicobacter pylori, Mycoplasma genitalium, Methanococcus jannaschii, Mycoplasma pneumoniae, Methanobacterium thermoautotrophicum, Mycobacterium tuberculosis, Pyrococcus abyssi, Rickettsia prowazekii, Synechocystis sp., Thermotoga maritima, Treponema pallidum, & Saccharomyces cerevisiae* (chromosome 4), which were retrieved from the National Center of Bioinformatics World Wide Web server (http://ncbi.nlm.nih.gov/). To avoid a bias due to each particular codon usage preference, the retrotranslation process was done using a unique codon usage table (from *E. coli*), retrieved from the Codon Usage Database (http//www.kazusa.org.jp/codon). The codon usage data was rewritten as a weighed probability matrix and used to assign codons to each amino acid residue coming from the proteome sequences. The curvature profiles of the resulting DNA sequences were calculated using the algorithm of Goodsell and Dickerson (1994). For each nucelotide of the sequence, a curvature value is assigned, expressed as a deviation angle from the helical axis per helix turn. This value is based in the contribution matrix for rotational and spatial displacements reported by Satchwell et al. (1986), because it has been used to predict accurately the curvature of well-characterized curved DNA segments. This matrix assigns each triplet a value of twist, roll, and tilt; these topological variables describe the position of a base pair with respect to its neighbors. The twist value is assumed constant (34.3 degrees) and the tilt value is assumed as 0; therefore, the roll angle values are the only source of variation of the curvature patterns. Average values from a 31 bp sliding window are collected at each nucleotide position and presented as normalized cumulative-frequency histograms (Fig. 1). This window size has been proved to be the most accurate to reproduce experimental results (Goodsell and Dickerson, 1994).

The difference between profiles coming from real and permuted proteomes was evaluated as follows. Permuted proteomes were generated by repositioning each residue at random within the sequence. The retrotranslation process was repeated 1000 times for both natural and permuted versions, and the means of the curvature values were collected to obtain normal distributions. The distance, in standard deviation units, between the averages of the normal distributions of real and permuted proteomes, permitted us to assess the statistical significance of the differences between the curvature average values from natural retrotranslated proteomes and their permuted versions, as described in Jáuregui et al. (1998). A similar procedure was used to evaluate the statistical significance of the differences arising from the comparison of individual proteins. All programs were written in Perl and C programming languages. (The source code is available at http://www.ibt.unam.mx/~ruy/programs/).

A databank of the orthologous elements common to *H. pylori, M. tuberculosis,* and *E. coli* was obtained using the Smith and Waterman global alignment algorithm (1981), provided by the Fasta Package (Pearson, 1991), using an arbitrary cutoff value of $p < 10^{-6}$.

RESULTS

Curvature of DNA obtained from retrotranslated proteomes

For every available proteome, an artificial DNA sequence was generated by retrotranslation and the corresponding curvature profiles were obtained. Differences between profiles were evident, each retrotranslated

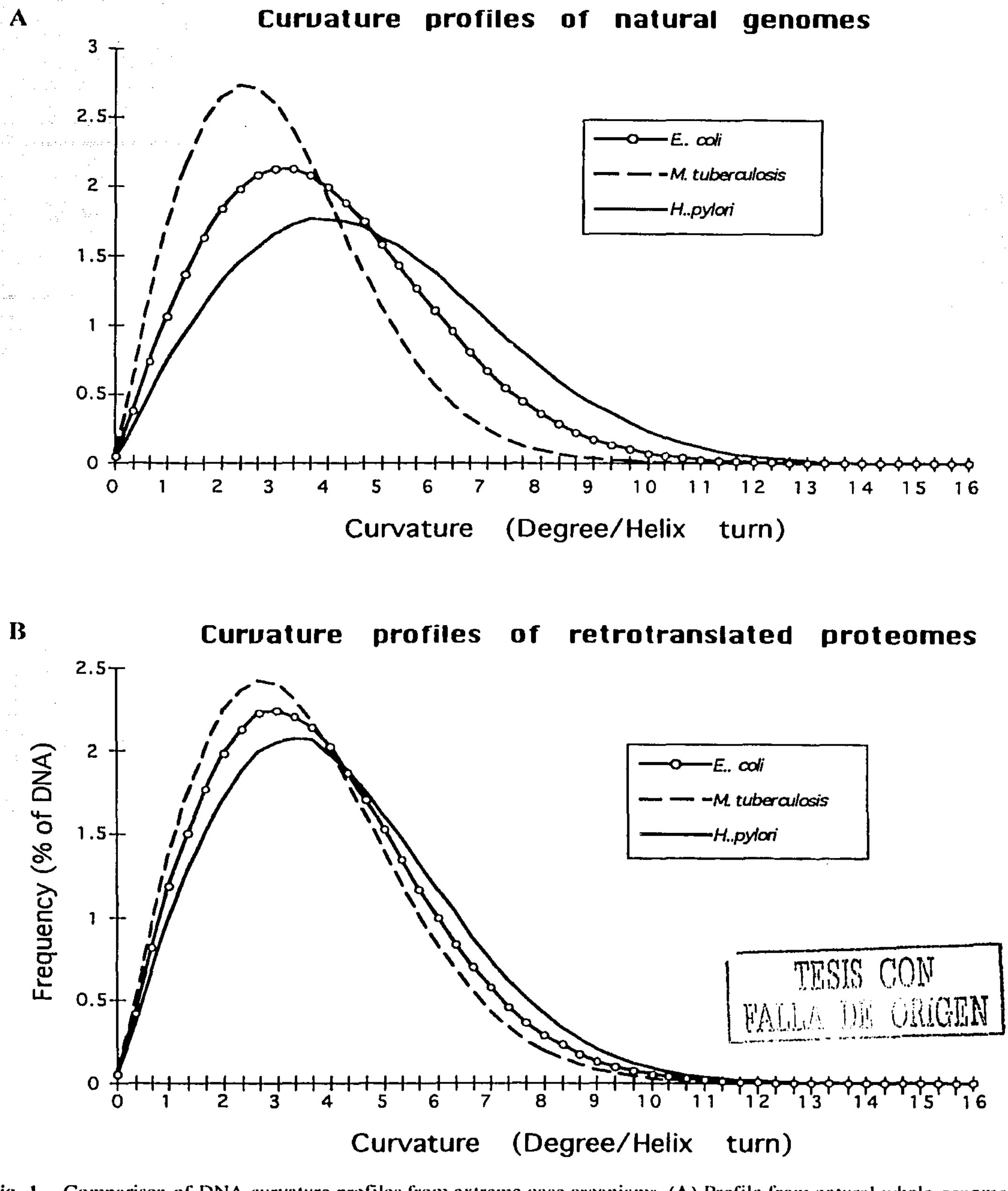


Fig. 1. Comparison of DNA curvature profiles from extreme case organisms. (A) Profile from natural whole genome DNA sequences of organisms that presented the lowest (*M. tuberculosis*, dashed line), medium (*E. coli*, line with circles), and highest (*H. pylori*, solid line) DNA curvature averages. (B) Profiles from synthetic DNA obtained by the retrotranslation of the proteomes from the organisms mentioned above. Only one codon usage table, from *E. coli*, was used to avoid the bias that codon preferences imprint on DNA curvature.

proteome following the tendencies of their natural genomes, indicating a relationship between the genomic DNA curvature of an organism and its proteome sequence. Representative cases are shown in Figure 1.

To discard the possibility that the differences in the curvature profiles could be originated by a particular set of proteins, we made a databank of orthologous elements common to the organisms that presented the highest *(H. pylori)*, lowest *(M. tuberculosis)*, and medium *E. coli)* DNA curvature averages. According to the criteria mentioned in Materials and Methods, 423 proteins were selected, and the previous analysis was repeated with this data set. The profiles obtained using this ortologous databank were almost identical to the ones obtained from the whole-proteome analysis, indicating that the contribution to the curvature observed in the analysis of whole proteomes is not due to a specific group of proteins, but is rather a general characteristic of them (data not shown).

To determine whether the behavior observed for proteomes was evident even in a single protein, we repeated a similar retrotranslation analysis using the peptidic sequences of the DNA Pol. III enzyme from *H. pylori* (ID g2314638) and *M. tuberculosis* (ID g1403495), which share 30% of identical residues. We obtained the average curvature values of 1000 retrotranslation cycles, and grouped them into cumulative frequency histograms. Normal distributions were evident and their mean and standard deviation were calculated. Statistically important differences were observed, as the means of these two profiles were more than 5 standard deviations (SDs) apart. This analysis was extended to every *H. pylori* and *M. tuberculosis* orthologous pair. In 97% of the cases the means of the normal distributions were more than 3 SDs apart (data not shown).

A comparison between the normalized frequencies of aminoacid usage for every full-sequenced organism was made (Table 1). In this table we also included the values corresponding to curvature averages from synthetic DNA coming from the retrotranslation of the proteomes. These values were used to sort the organisms in the table and to find a possible correlation with aminoacid frequencies. To make fair comparisons, we used a standard codon table, from *E. coli*. Clear cases of high correlation and anticorrelation were found, such as Lysine (0.95) and Proline (−0.86), respectively. The implications of these results are discussed later.

Curvature of DNA obtained from permuted and retrotranslated proteomes

To determine whether the behaviour described above is related only to the overall aminoacid composition of the proteome or could be also a position-dependent effect, we compared the curvature profiles from the retrotranslated sequences of a natural proteome and its permuted version. The overall curvature profiles were very similar; the average curvature differences were under 0.1 degrees per helix turn, indicating that the differences observed between curvature profiles of retrotranslated proteomes are mainly related to aminoacid composition.

Considering that the whole proteome universe is being sampled, and a curvature value is assigned to each nucleotide of the sequence, even the small differences observed in curvature averages coming from natural and permuted proteomes might prove to be significant. To evaluate the position-dependent contribution, we compared the normal distributions of curvature averages arising from natural and permuted sequences after 1000 retrotranslation cycles of *E. coli* and *H. pylori* proteomes. The resulting distributions demonstrated a statistically significant position-related contribution because their mean values were 5 and 8.4 SD apart, respectively (Fig. 2). To test the extent of this phenomenon, for each individual protein of every fully sequenced organism, we compared the average curvature values arising from real and permuted peptidic sequences. For the majority of the proteins, in almost every organism, the permutation has a negative effect on the curvature values of the retrotranslated sequence (Fig. 3). The most extreme case was *H. pylori*, where 75% of the retrotranslated proteins presented a higher curvature average than their permuted sequences. It is worth noting that this organism also presented the highest natural DNA curvature average. On the other hand, exceptions to this tendency are observed in *Bacillus subtilis* and *Chlamidia pneumoniae*, where in the majority of their proteins, the permuted version induces a higher average DNA curvature than its original peptide sequence.

DISCUSSION

Although the function of *loci*-specific DNA curvature in small regions has been well understood and examined (reviewed by Travers, 1990; Hagerman 1990; Harrington, 1992; Kathleen, 1992), the biological im-

Table 1. Normalized Aminoacid Frequencies and Curvature Averages of Retrotranslated Proteomes

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*
cobacter pylori	6.83	1.09	4.77	6.88	5.41	5.76	2.12	7.2	8.94	11.1	2.28	5.83	3.28	3.71	3.46	6.82	4.37	5.59	0.7	3.6
anococcus jannaschi	5.51	1.28	5.5	8.63	4.24	6.39	1.43	10.5	10.3	9.41	2.32	5.25	3.37	1.43	3.84	4.48	4.06	6.84	0.71	4.3
lla burgdorferi	4.49	0.66	5.19	6.77	6.31	5.2	1.22	10.7	10.2	10.3	1.9	7.27	2.52	2.26	3.22	7.47	3.94	5.36	0.5	4.2
plasma genitalium	5.57	0.83	4.9	5.64	6.12	4.63	1.57	8.25	9.47	10.6	1.53	7.52	3	4.73	3.09	6.65	5.4	6.11	0.97	3.2
ettsia prowazekii	6.04	1.09	4.83	5.78	4.88	5.41	1.9	10.8	8.37	10	2.17	6.65	3.14	3.14	3.39	6.75	5.21	5.59	0.71	3.8
fex aeolicus	5.89	0.79	4.31	9.63	5.14	6.75	1.54	7.33	9.4	10.5	1.93	3.59	4.07	2.04	4.92	4.79	4.21	7.93	0.93	4.1
eoglobus fulgidus	7.85	1.17	4.89	8.9	4.59	7.24	1.51	7.22	6.86	9.5	2.62	3.21	3.86	1.78	5.77	5.51	4.16	8.61	1.03	3.6
coccus abyssi	6.67	0.55	4.6	8.84	4.35	7.26	1.5	8.49	7.8	10.2	2.4	3.33	4.25	1.66	5.72	4.97	4.2	8.07	1.17	3.8
coccus horikoshii	6.37	0.63	4.26	8.29	4.6	6.97	1.49	8.78	7.74	10.3	2.4	3.53	4.5	1.63	5.45	5.85	4.51	7.55	1.17	3.8
notoga maritima	5.85	0.7	4.96	8.92	5.19	6.9	1.58	7.18	7.61	10	2.4	3.61	3.99	2.01	5.53	5.65	4.52	8.61	1.1	3.5
plasma pneumoniae	6.66	0.74	4.97	5.68	5.59	5.52	1.8	6.6	8.56	10.3	1.58	6.22	3.49	5.37	3.48	6.46	5.96	6.47	1.18	3.2
nophilus influenza	8.21	1.03	4.97	6.48	4.47	6.64	2.05	7.1	6.31	10.5	2.44	4.87	3.71	4.64	4.47	5.84	5.2	6.67	1.13	3.1
llus subtilis	7.67	0.8	5.17	7.23	4.49	6.9	2.27	7.36	7.05	9.64	2.78	3.94	3.68	3.83	4.12	6.29	5.42	6.74	1.03	3.4
haromyces cerevisiae	5.38	1.3	5.94	6.71	4.37	4.88	2.17	6.58	7.48	9.57	2.05	6.26	4.26	3.84	4.49	9.05	5.8	5.47	0.98	3.3
orhabditis elegans	6.19	1.99	5.39	6.56	4.8	5.27	2.33	6.11	6.46	8.7	2.58	4.93	4.82	4.09	5.29	8.11	5.87	6.23	1.1	3.0
mydia trachomatis	6.98	1.59	4.51	6.61	4.74	6.23	2.38	6.92	6.14	11.3	1.93	3.81	4.48	4.03	4.56	8.02	5.27	6.12	1.01	3.2
mydia pneumoniae	7.52	1.62	4.52	6.6	4.83	6.34	2.3	6.6	5.75	11.2	2.05	3.5	4.36	4.18	4.84	8.12	5.11	6.42	0.95	3.0
erichia coli	9.48	1.17	5.13	5.74	3.89	7.36	2..27	6	4.4	10.6	2.85	3.95	4.42	4.42	5.53	5.82	5.4	7.05	1.52	2.8
hermoautotrophicum	7.33	1.2	5.91	8.14	3.64	7.97	1.87	7.7	4.56	9.42	3.07	3.31	4.3	1.9	6.79	6.14	4.96	7.66	0.84	3.2
chosisitis sp.	8.49	1	5.02	6.04	4.01	7.37	1.86	6.28	4.18	11.4	2.01	4.04	5.14	5.55	5.05	5.81	5.5	6.69	1.55	2.9
onema pallidum	10.1	1.91	4.52	5.97	4.45	6.96	2.75	4.9	3.97	10.1	2.09	2.48	4.2	3.84	7.43	6.62	5.3	8.25	0.97	3.0
phyrum pernix	9.51	0.94	3.87	6.61	2.74	8.55	1.92	5.19	3.54	11.3	2.21	2.03	6.45	1.9	7.72	7.52	4.68	8.5	1.31	3.3
bacterium tuberculosis	13.2	0.88	5.8	4.67	2.95	9.98	2.23	4.26	2.03	9.74	1.95	2.53	5.8	3.09	7.32	5.48	5.92	8.46	1.47	2.0
elation index	−0.82	−0.3	−0.07	0.45	0.77	−0.68	−0.62	0.81	0.95	−0.09	−0.09	0.66	−0.86	−0.2	−0.78	−0.2	−0.56	−0.43	−0.66	0.8

r each fully sequenced organism, the relative abundance of each aminoacid was evaluated and expressed as a normalized frequency. Curvature averages of the retrotranslated proteome usi
n preferences are included and used to order the organisms. The correlation values between aminoacid frequency and DNA curvature are indicated at the bottom of the table, and the mo
nns are shaded.

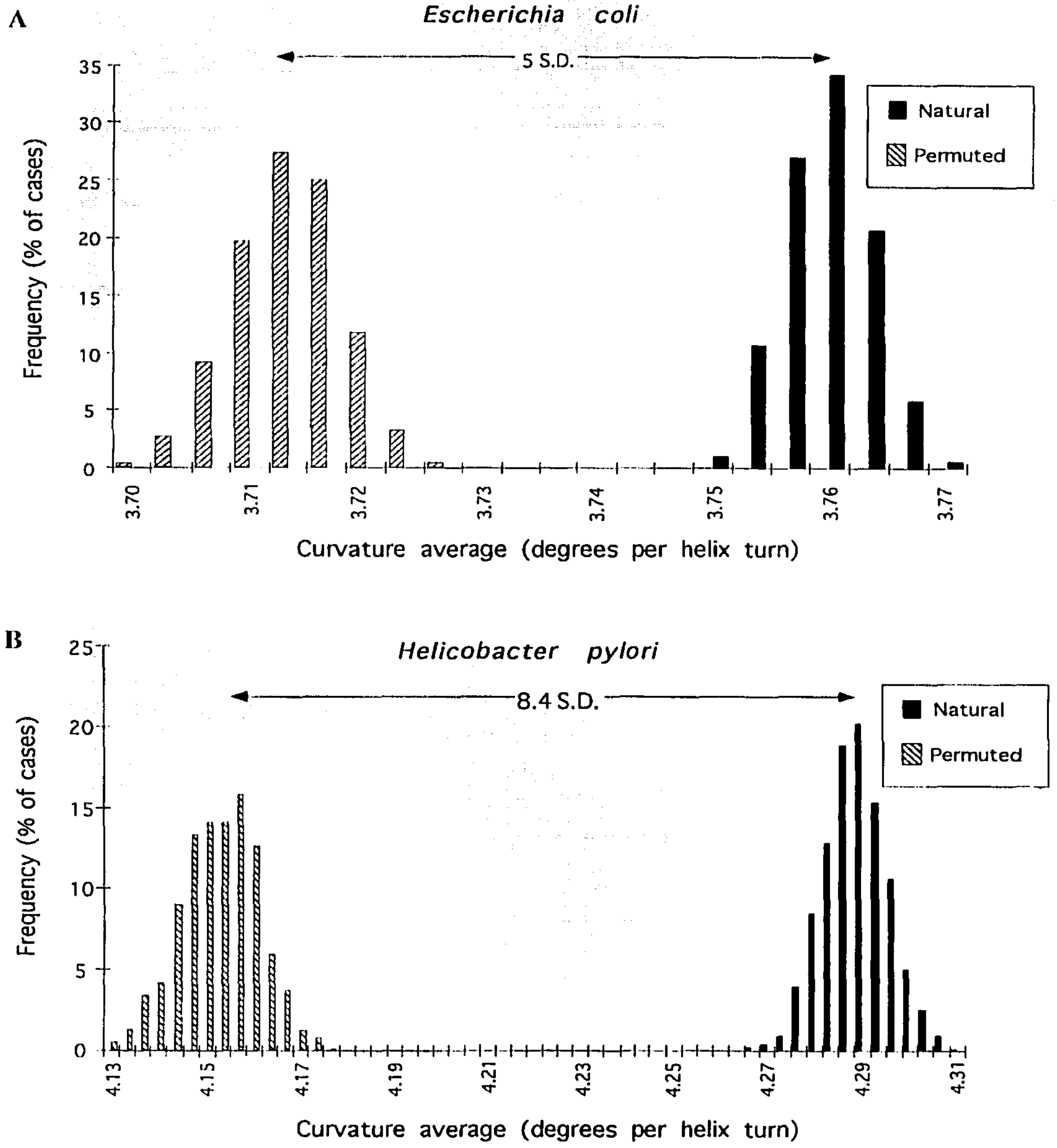
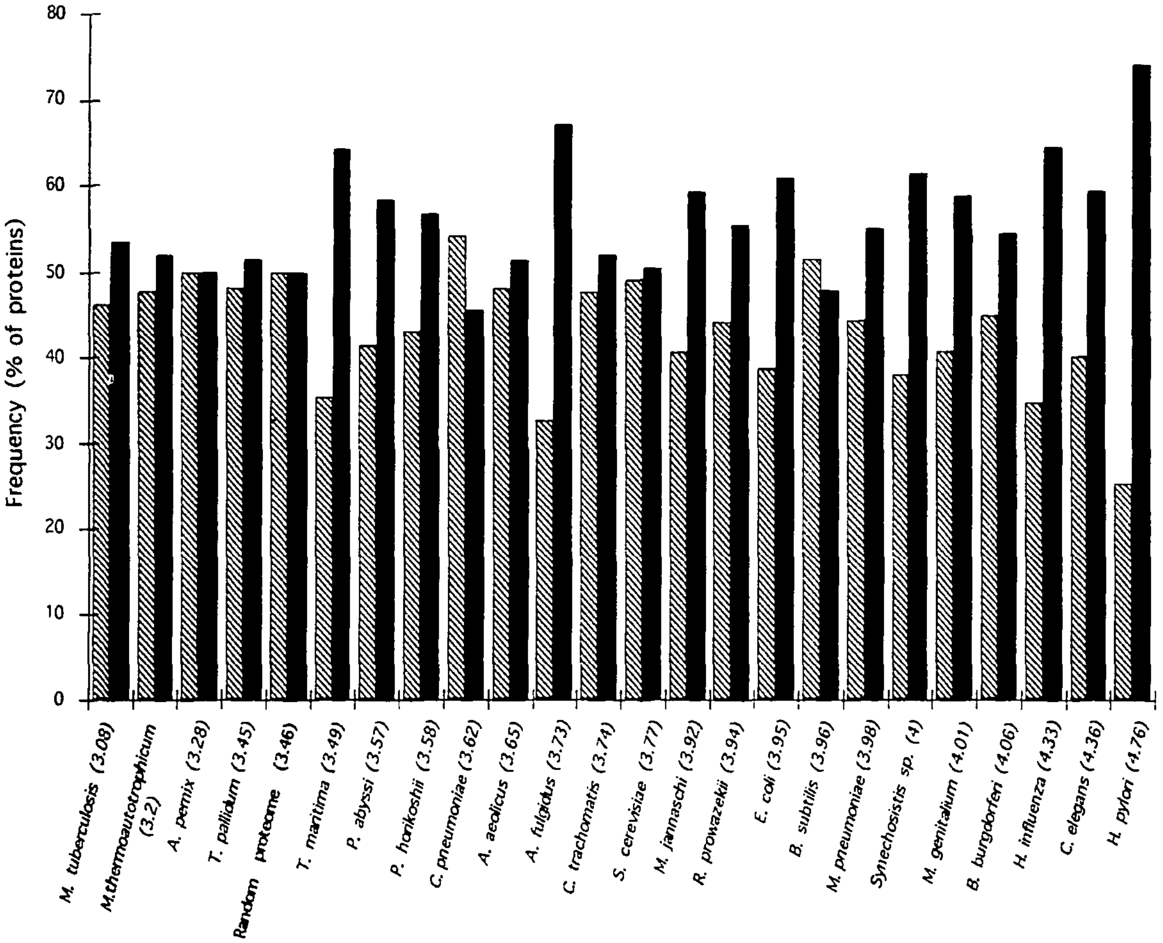


Fig. 2. Normal distributions of DNA curvature averages from natural (solid bars) and permuted (hatched bars) retro-translated proteomes. The aminoacid sequences of organisms with (**A**) medium *(E. coli)* and (**B**) high *(H. pylori)* DNA curvature values were analyzed. Curvature averages were obtained by repeating 1000 times the processes of retro-translation and DNA curvature calculation. Distances between the normal distribution means are indicated in standard deviation units (SD).



3. Comparisons of DNA curvature averages from retrotranslated aminoacid sequences, natural and permuted. Every single protein was retrotranslated 1000 times and
ature of its corresponding DNA sequences was evaluated. The average of these curvature values were obtained and compared with the corresponding average derived fron
nuted aminoacid sequence. Bars indicate the percentage of the cases where the natural aminoacid version presented a higher (black) or lower (dashed) curvature average t
ermuted version. The organisms are sorted according to its natural genome curvature average, indicated next to the organisms' names. The distribution of a synthetic protec
an equimolar composition of aminoacids in random order, and retrotranslated with an equiprobable codon table, is also included as a reference.

plications of DNA curvature considering the complete genome of an organism have not been addressed until recently. The nonfortuitous nature of whole-genome curvature profiles has been demonstrated, and characteristic organism-specific curvature profiles have been described (Gabrielian et al., 1997; Jáuregui et al., 1998). In our previous study, the first 9 full-sequenced bacterial genomes available (at the TIGR database www.tigr.org) presented higher curvature values than the ones obtained from random sequences, leading us to the hypothesis that the DNA curvature could be an important factor in bacterial chromosome condensation due to the lack of efficient DNA packing proteins, such as histones. The analysis of recently sequenced bacterial genomes has invalidated this hypothesis, as many new organisms presented lower DNA curvature values than their randomized genomic sequences.

Even when the biological meaning of whole genome DNA curvature is still in debate, we have found an important relationship between the organism's DNA curvature profile and its codon usage (Jáuregui et al., 1998). In addition, the influence of dinucleotide frequencies of genomes in DNA cuvature is under study (Merino and Garciarrubio, 2000).

Genomic sequence analysis has revealed biases in the aminoacid composition of proteomes in different organisms (see Table 1). The biological implications of these biases are not well understood, and only a few cases provide a clear explanation for them, as in the *M. tuberculosis* proteome; that is, biased due to the high GC content of its genome (Cole et al., 1998).

Here we address the question of whether the bias in the aminoacid composition of proteomes could have an impact on genomic DNA curvature. To discriminate the particular contributions of the proteome from other sources of variation, such as the codon usage preferences, we devised a program that retrotranslated the aminoacid sequence of different proteomes into DNA, using a unique codon preference table for all of them.

The retrotranslation process yielded *in silico* DNA sequences that are biased toward its natural genome curvature, demonstrating a relationship between the aminoacid composition of a proteome and the DNA curvature of its genome. It is important to stress that the differences in the curvature profiles of the artificial DNA sequences are an exclusive proteome-encoded contribution. It is remarkable that this phenomenon can be observed even in the comparison of a single homologous protein pair. Differences in homologue protein sequences can originate changes in the protein structure and/or function, or can be regarded as neutral (Jukes and Kimura, 1984); nevertheless, here we have presented evidence that these changes, though neutral in the polypeptide sequence, can influence the DNA curvature profile.

The retrotranslation analysis of natural and randomly permuted proteomes showed that the main determinant of proteome-encoded DNA curvature is the aminoacid composition. Although this result, in principle, could be taken as a straightforward inference, because the aminoacids with high curvature triplets might be thought to be the principal contributors to overall curvature profiles, it was found that it was not the case. The inspection of the aminoacid frequencies of proteomes and the corresponding curvature values of their derived DNA sequences revealed that the most important contribution came from Lys, Pro, Ala, Ile, Arg, and Phe residues, as indicated by the correlation index of Table 1. In the case of Lysine, the aminoacid with the highest correlation index (0.95), its triplets AAA and AAG have small roll angles of 0 and 4.2 degrees, respectively. In the same trend, the triplet with the greatest roll angle (8.1) is GCC, coding for Ala, which is strongly anticorrelated with curvature (its correlation index is −0.82). A detailed examination of the Goodsell and Dickerson algorithm (1994) used in our study revealed that the magnitude of the DNA curvature depends on the variations and phase of the roll values. A single triplet with a constant roll value, as high as it might be, in a repetitive sequence would produce no curvature because the contribution would annihilate itself in each complete helix turn, due to the fact that in an ~5 nt segment the rotation of the bases places the roll angle in the opposite side of the helix, which amounts to add the negative value of itself, giving an average contribution of zero. In the other hand, a triplet with a relatively small roll value occurring in phase (every ~10 nt.) within a noncurved sequence, would add to the roll values and produce a high curvature contribution.

The comparison between curvature profiles arising from the retrotranslation of natural and permuted proteomes revealed a statistically significant contribution of the aminoacid order within the peptide sequence to the overall genomic DNA curvature. Even when the absolute curvature averages difference is small, the large size of the universe examined allowed us to determine that this difference is indeed significant and must be taken into consideration. It is clear that the main contribution of the proteome to DNA curvature is due to its aminoacid content. Nevertheless, the relative order of the aminoacids is also important, because the out-of-phase triplets starting at the second and third base of a codon also influence DNA curvature. In this regard, the comparisons of curvature averages arising from the retrotranslation of natural and permuted proteins give us a remarkable result: the aminoacid position-dependent curvature contribution in almost all the organisms is positive (20 out of 23). This data points toward the existence of a natural selection of genomic DNA curvature. Further investigation might shed light on the relevance of global curvature of genomes.

ACKNOWLEDGMENTS

We thank Guillermo Gosset, Gabriel Moreno-Hagelsieb, Miguel Angel Cevallos, and Eduardo Horjales for their comments on the manuscript; Ricardo Ciria, Abel Linares, Juan Manuel Hurtado, and Alma Martinez for computer support; and Shirley Ainsworth for bibliographic assistance. This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT)-UNAM, Grant IN230798.

REFERENCES

COLE, S.T., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393–11**, 537–544.

DIEKMANN, S. (1989). Definitions and nomenclature of nucleic acid structure parameters. EMBO J **8–1**, 1–4.

GABRIELIAN, A., VLAHOVICEK, K., and PONGOR, S. (1997). Distribution of sequence-dependent curvature in genomic DNA sequences. FEBS Lett **406**, 69–74.

GOODSELL, D.S., and DICKERSON, R.E. (1994). Bending and curvature calculations in B-DNA. Nucl Acid Res **22**, 5497–5503.

HARRINGTON, R.E. (1992). DNA curving and bending in protein-DNA recognition. Mol Microbiol **6**, 2549–2555.

HAGERMAN, P.J. (1990). Sequence-directed curvature of DNA. Ann Rev Biochem **59**, 755–781.

JÁUREGUI, R., O'REILLY, F., BOLIVAR, F., and MERINO, E. (1998). Relationship between codon usage and sequence dependent curvature of genomes. Mic & Comp Genom **3–4**, 243–253.

JUKES, T.H., and KIMURA, M. (1984). Evolutionary constraints and the neutral theory. J Mol Evol **21**, 90–92.

KARLIN, S., MRÁZEK, J., and CAMPBELL, A.M. (1997). Compositional biases in bacterial genomes and evolutionary implications. J Bacter **179–12**, 3899–3913.

KATHLEEN, S.M. (1992). DNA looping. Microbiol Rev **56**, 123–136.

MERINO, E., and GARCIARRUBIO, A. (2000). The global intrinsic curvature of archaeal and eubacterial genomes is mostly contained in their dinucleotide composition and is probably not an adaptation. Nucl Acids Res **28**, 2431–2438.

MRÁZEK, J., and KARLIN, S. (1998). Compositional asymmetry in bacterial and large viral genomes. Proc Nat Acad Sci USA **95**, 3720–3725.

PEARSON, W.R. (1991). Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics **11**, 635–650.

PÉREZ-MARTIN, J., ROJO, F., and De LORENZO, V. (1994). Promoters responsive to DNA bending: A common theme in prokaryotic gene expression. Microbiol Rev **58**, 268–290.

SATCHWELL, S.C., DREW, H.R., and TRAVERS, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. J Mol Biol **191**, 659–675.

SMITH, T.F., and WATERMAN, M.S. (1981). Identification of common molecular subsequences. J Mol Biol **147**, 195–197.

TRAVERS, A.A. (1990). Why bend DNA? Cell **60**, 177–180.

WATSON, J.D., and CRICK, F.H.C. (1953). Molecular structure of nucleic acids. Nature **171**, 737.

Address reprint requests to:
Dr. Enrique Merino
Instituto de Boitecnologia, UNAM
Av. Universidad 2001, Chamilpa
c.p. 62210, Cuernavaca, Morelos, México

E-mail: merino@ibt.unam.mx

Conservation of DNA Curvature Signals in Regulatory Regions

Ruy Jáuregui*, Gabriel Moreno-Hagelsieb#, Cei Abreu Goodger*, Julio Collado-Vides#, and Enrique Merino*

** Instituto de Biotecnologia, Universidad Nacional Autónoma de México, Av. Universidad 2001, Chamilpa,Cuernavaca Mor., México. # Program of Computational Genomics, CIFN-UNAM, Apdo Postal 565-A, Cuernavaca, Morelos, 62100 Mexico*

Abstract

DNA curvature plays a well-characterized role in many transcriptional regulation mechanisms. In this work we present evidence for the conservation of curvature signals in putative regulatory regions of several genomes. Genes with highly curved upstream regions were collected into orthologous groups, based on the annotations of the Cluster of Orthologous Groups (COGs) database. COGs with a significant number of genes with curvature signals were analyzed and conserved properties were identified in several cases. We found curvature-mediated regulatory sites, previously described in single organisms, in a broad spectrum of bacterial genomes, stressing the fact that structural parameters of the DNA molecule, beyond the sequence, are conserved elements in the process of transcriptional regulation.

Introduction

Studies of the relationship between DNA curvature and transcription regulation have been conducted mostly for specific sets of genes and discrete *loci.* Experimental evidence has demonstrated a contribution of DNA curvature in regulating the transcription of several genes, such as H-NS histone-like protein (30), Sigma-S (12), IHF and HU regulatory proteins (7), Sigma-54 dependent *glnAp2* and *glnHp2* (32), and artificial constructs using the T7 virus promoter (1 1), among others.

The first genome-wide analysis of promoter sequences in *E coli*, found these regions to be significantly more curved than coding regions or randomly permuted sequences. (10) This data is consistent with our previous observations in whole-genome non-coding regions (8). Moreover, binding sites for known regulatory proteins were found to present even higher curvature values (10). It was also proposed that mesophilic bacteria, as opposed to hyperthermophilic bacteria and archaea, present a bias to a high DNA curvature content due to a temperature dependent transcription regulation (43). In these previous studies, only average curvature values were considered, therefore individual genes with significant curvature signals were not identified.

As far as we know, no attempt has been made to establish if discrete DNA curvature signals are conserved regulation features in different organisms. Here we extend these previous studies and address the question of whether static DNA curvature is a conserved feature of the transcriptional regulation mechanisms within the broad context of 98 available microbial genomes (see Materials and Methods). Using the data compiled in the COG (Cluster of Orthologous Genes) database (1), and additional orthology data for the genomes not included in it (see Materials and Methods), we demonstrate a significant conservation of curvature signals within the regulatory regions of several orthologous gene clusters spanning a broad spectrum of biological functions,. Among these set of orthologous genes, DNA-binding proteins were found to present curvature signals in a large number of genomes. A detailed examination of the most relevant cases is presented.

Materials and Methods

DNA sequence data.

DNA sequence was derived from the complete bacterial genomes available in the Entrez Genome Database, (ftp://ncbi.nlm.nih.gov/genomes/Bacteria/). Different strains of the same organism were eliminated, leaving the one with the largest genome. Thus, the analysis contemplated 98 complete archaeal and bacterial genomes. A list of these genomes can be found at (www.ibt.unam.mx/bioinfo/curvature-genomes)

Delimitation of regulatory regions.

A 250 nt. window containing 200 bases upstream and 50 bases downstream of the start codon of each coding sequence (CDS) was chosen as our analysis window, since more than 90% of the regulatory signals are found within this range in *E. coli* K12 (58). The set of upstream regions for each organism was obtained taking into account the operon organization of its genes. Operon prediction was based on inter-genic distances as described by Salgado and Moreno-Hagelsieb (2, 45). The upstream region associated to a gene is the region upstream of the first gene in its operon and is defined as the Minimal Upstream Region (MUR).

Curvature calculations.

DNA curvature was calculated using the computer algorithm BEND (6) and the rotational and translational contribution matrix derived from nucleosome positioning sequence data (46). A curvature profile is obtained by assigning each nucleotide of the sequence a curvature value, expressed as a deviation angle from the helical axis per helical turn. Signal-to-noise ratio was minimized by taking the average value of a sliding window of 31 nucleotides (approx. 3 helical turns), and assigning it to the central nucleotide.

Since each genome presents a distinctive curvature profile (14, 8), curvature average and standard deviation (SD) values were obtained for every genome considered in this study. A cutoff value of 3 SD from the genomic curvature mean was used to identify statistically significant signals in the set of MURs, and their genes were collected and sorted into their corresponding orthologous groups.

Clustering of orthologs.

Our orthologous gene sets were mainly those found in the COGs database (1). Genomes not included in this database were compared to annotated genomes using gapped-BLAST (35). Orthologs were identified and added to the corresponding COG using the bi-directional best hit criterion, adding the requirement that at least 50% of the smallest sequence was included in the alignment.

Identification of COGs with statistically significant number of curvature signals.

To evaluate the statistical significance of the number of genes with curved DNA signals in their regulatory regions in a given COG, we used the following procedure: a) we generated a database of Montecarlo permutations of the MURs within each complete genome. b) we counted the number of genes within each COG that presented a curvature signal in its MUR. c) Steps a) and b) were repeated 1000 times to find the mean and standard deviation of the number of signals for each COG. d) We estimated the statistical significance of the number of signals in the COGs from the real genomes by measuring its distance, in standard deviations, to the mean of the values obtained from the Montecarlo permutations. (tables 1 and 2). COGs with less than 5 organisms represented were excluded from our analysis.

Promoter prediction.

Promoter sequences for the genes in significant COGs were predicted using the algorithm of Mulligan et al., (47). Weight matrices were derived from alignments of experimentally characterized promoters for sigma 70 (47) and sigma 54 (48) promoter sequences. Regions containing the best scoring promoters plotted using the DIAMOD DNA curvature display software (49).

Results.

Analysis of COGs with conserved curvature signals.

Sixty COGs presented a statistically significant number of curvature signals (above 3 SD from the mean obtained in the randomization procedure). These COGs were classified accordingly to its given global functional characterization (1) (Table 1). Experimental data to support DNA curvature related to transcriptional regulation for these cases was searched for in the literature. Biologically relevant COGs with lower scores (over 2 DS) were also included (table 2) and considered in our analysis. Representative cases of the best scoring COGs are:

i. Proteins HU and IHF from COG0776. 64 genes from 43 different organisms were found to have curvature signals in their regulatory regions. Both IHF and HU proteins are known to be key regulators in a broad spectrum of genes in several organisms. These proteins bind to curved DNA regions and further bend the DNA molecule (23). Autonomous transcriptional regulation for the *hupA* and *hupB* genes in *E. coli* has been demonstrated (41) . Besides this auto-regulation, the transcription of these genes has been found to be dependent on CRP (Catabolite Repression Protein) and FIS (Factor for Inversion Stimulation) regulators, both of which bind to curved DNA (24). In the case of the genes coding for the IHF dimmer, *himA* and *himD*, auto regulation has also been demonstrated along with dependence of rpoS and ppGpp levels (25, 26, 50).

ii. DNA gyrase subunits A and B from COG0188 and COG0187, respectively. 45 genes for subunit A were found in 37 organisms, and 45 genes for subunit B in 38 organisms. DNA gyrase is responsible for negatively supercoiling the DNA molecule, and its own transcription has been demonstrated to be regulated by the modulation of the supercoiling state of the DNA molecule (51, 52). A bent DNA region between the -35 and -10 elements of the GyrA promoter in *S. pneumoniae*, has been described, and it has been proposed that this region makes the promoter very sensitive to changes in supercoiling, allowing the expression of GyrA to act as a regulator of DNA supercoiling in the cell (42). Significant curvature signals have also been predicted in several micobacterial gyrase promoters (44).

iii. Aspartyl tRNA synthetase from COG0173, with 25 genes in 25 organisms. The presence of conserved Upstream Activating Sequences (UAS), regulating bi-directional promoters of glutamyl-tRNA synthetase and the *valU* and *alaW* tRNA operon in *E. coli* has been demonstrated (17). The DNA in the UAS is known to be bent (19, 20) and also be a target for the FIS regulatory protein (21).

iv. Transposase from COG3385, with 38 genes in 6 organisms. Transpositional modulation has been found to be dependent on global regulatory proteins such as HU, IHF and H-NS (37, 38, 39) all of them known to bind to curved DNA. Interestingly, the role of DNA curvature involved in transposition has been confirmed for the insertion sequence IS231A in *Bacillus thuringiensis,* where one of the terminal repeats of the transposon Tn4430 was found to be an insertional hot spot, due to the flanking curved DNA regions (15). Even though this evidence is not directly involved in transcriptional regulation, our data support the idea of a conserved curvature profile related to transposition, and confirm a general role of DNA curvature in transposition events. The presence of curvature signals in the transposase regulatory region suggests a role of DNA curvature in transcriptional regulation; nevertheless, this has still to be verified.

v. 30S ribosomal protein S20 from COG0268. 23 genes from 23 organisms presented curvature signals. Although no direct evidence has been reported for the relevance of DNA curvature in the transcription of the gene that codes for this protein, the role of FIS regulated UAS has been documented for several ribosomal operons (40, 54).

vi.Cell division related genes from COG3116 with 11 genes from 11 organisms (11/11), COG0552 with 29 genes from 29 organisms (29/29), COG3096 (5/5), COG3006 (5/5), COG0849 (21/21), and COG3095 (5/5) . In this case we found different COGs involved with the genome replication and cell division process. The time coordination requirements for such process impose a highly regulated transcription schedule. This transcription is in many cases mediated by general DNA curvature dependent regulators such as IHF, FIS and HU (55, 56).

vii.Glutamine synthetase from COG0174, with 46 genes from 31 organisms. This gene, with a sigma 54-dependent promoter, has been found to be also dependent on a bent region between the promoter and the enhancer site to initiate transcription in *E. coli.*

(32, 43). Our finding of a conserved curvature signal in the regulatory regions of these genes in 31 genomes indicates that this mechanism is widely conserved. In an attempt to further characterize the regulatory regions of these genes we predicted sigma 54 dependent promoters, using a weight matrix derived from the compilation of 186 sequences reported in the literature (44) and are presented in Figure 3).

Visualization and analysis of conserved curvature signals.

A detailed examination of the DNA curvature profiles of the MURs from the significant COGs was conducted in order to identify common and conserved properties. Representative examples of COGs

COGs grouping HU and IHF orthologs, DNA gyrase and FIS were plotted using the DIAMOD DNA curvature display software (49), promoter prediction showed several cases were the maximal curvature value was less than 30 bp from the -35 element of the promoter sequence. Fig 1 presents the curvature profiles of predicted promoter regions for these COGs.

The curvature profiles of regulatory regions of ribosomal RNA genes, known to be regulated by FIS and where promoter sequences had been experimentally determined, demonstrated conserved curvature signals contiguous to the promoter, even among sequences with low sequence homology, curvature profiles of ribosomal protein S20 from several organisms, also dependent on FIS, demonstrated conserved curvature signals near the predicted promoter region (fig2).

The sigma 54 dependent glutamine synthetase gene *glnA*, whose promoter region is known to be bent in *E. coli* (57), was found grouped with orthologs from 31 other genomes with curvature signals, a detailed analysis demonstrated conserved curvature signals near predicted sigma 54 promoters (fig3).

There are other ribosomal genes who are also regulated by curvatura (fig 2).

Discussion.

The structure of the DNA molecule, in this case DNA curvature, has been found to play important roles in several biological processes, including DNA replication and packaging, chromosome segregation, recombination, transposition, virus integration and transcriptional regulation. Until recently, DNA curvature had been studied in the context of discrete DNA fragments and particular loci in single organisms, and no attempt had been made to find how general this mechanism was among different genomes. In this work we present a first attempt to study the conservation of DNA curvature as an element of transcriptional regulation, by identifying curvature signals present in groups of orthologous genes.

Our finding of a significant number of curvature signals in the upstream regions in several clusters of orthologous genes, added to the previous experimental characterization of a DNA curvature mediated regulatory mechanism in at least one of its members, provides evidence that curvature-mediated transcriptional regulation is widely conserved among several organisms.

The fact that global regulators such as HU and IHF present conserved curvature signals in their regulatory regions is expected, since they are known to bind to curved DNA, and to be autoregulated in at least one organism, the discovery of 40 bacterial genomes sharing a curvature motif suggests a highly conserved regulation mechanism.

Several gene groups related to cell division were an unexpected finding, since there is no experimental evidence of DNA curvature involved in their transcriptional regulation, but the presence of conserved curvature signals suggests a common regulatory protein or mechanism. Global morphological changes are known to occur to the chromosome during cell division and the idea of a conserved DNA structure playing a role in this process is not far-fetched.

Some previously unknown transcriptional regulators such as the araC family COG4977 were also found to be dependant on curvature, the detailed characterization of the structure of its binding sites might be helpful for the future detection of genes under the control of these proteins.

All these facts point towards a central role of the structure of the DNA molecule in transcriptional regulation. An integrated sequence and structure prediction approach for regulatory regions might result in more sensitive and efficient detection of regulatory motifs and a wider and clearer comprehension of regulation paradigms, since structural conservation is not necessarily sequence dependent.

Acknowledgements.

We thank Shirley Ainsworth for bibliographical support.

Figure legends.

Figure 1: DNA curvature plots of predicted promoter regions for DNA gyrase, HU and FIS orthologous genes. Promoter position is indicated by a black arrow.

Figure 2: Curvature profiles for experimentally characterized promoters for ribosomal RNAs in E. coli and predicted promoters for ribosomal protein S20 orthologous genes. Promoters are indicated by a black arrow.

Figure 3: Curvature profiles of promoter regions for glutamine sinthetase orthologs. The predicted promoter is indicated by a black arrow.

Table 1: Top significant COGs, the columns indicate the distance in Standard Deviation Units, the COG number, the number of genes/organisms, and the function according to the COG database classification.

Table 2: Biologically relevant COGs, which might present regulatory mechanisms mediated by DNA curvature are presented as an expansion of table 1, The columns indicate the Standard deviation distance units, the COG number, the number of genes/organisms and the associated function, as described in the COG database (1).

Bibliography.

1. Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin EV. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28, 33-36.

2. Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides. (2000) Operons in Escherichia coli: genomic analyses and predictions. J. Proc Natl Acad Sci U S A Jun 6;97(12):6652-6657

3. Perez-Martin J, de Lorenzo V. Clues and consequences of DNA bending in transcription. Annu Rev Microbiol 1997;51:593-628

4. Harrington RE. DNA curving and bending in protein-DNA recognition. Mol Microbiol 1992 Sep;6(18):2549-2555

5. HAGERMAN, P.J. (1990). Sequence-Directed curvature of DNA. Annu. Rev. Biochem. 59, 755-781.

6. GOODSELL, D.S., and DICKERSON, R.E. (1994). Bending and curvature calculations in B-DNA. Nucleic. Acid. Res. 22, 5497-5503.

7. PÉREZ-MARTIN, J., ROJO, F., and De LORENZO, V. (1994). Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. Microbiol. Rev. 58, 268-290.

8. JÁUREGUI, R., O'REILLY, F., BOLIVAR, F., AND MERINO, E. (1998). Relationship between codon usage and sequence dependent curvature of genomes. Mic & Comp. Genom. 3-4, 243-253.

9. Trifonov E.N. Curved DNA. CRC Crit Rev Biochem 1985;19(2):89-106

10. Gabrielian AE, Landsman D, Bolshoy A. Curved DNA in promoter sequences. In Silico Biol 1999-2000;1(4):183-96

11. Collis CM, Molloy PL, Both GW, Drew HR. Influence of the sequence-dependent flexure of DNA on transcription in E. coli. Nucleic Acids Res 1989 Nov 25;17(22):9447-68

12. Espinosa-Urgel M, Tormo A. Sigma s-dependent promoters in Escherichia coli are located in DNA regions with intrinsic curvature. Nucleic Acids Res 1993 Aug 11;21(16):3667-70

13. Shpigelman ES, Trifonov EN, Bolshoy A. CURVATURE: software for the analysis of curved DNA. Comput Appl Biosci 1993 Aug;9(4):435-40

14. GABRIELIAN, A., VLAHOVICEK, K., and PONGOR, S. (1997). Distribution of sequence-dependent curvature in genomic DNA sequences. FEBS letters 406, 69-74.

15. Hallet B, Rezsohazy R, Mahillon J, Delcour J. IS231A insertion specificity: consensus sequence and DNA bending at the target site. Mol Microbiol 1994 Oct;14(1):131-9

16. Figueroa N, Wills N, Bossi L. Common sequence determinants of the response of a prokaryotic promoter to DNA bending and supercoiling. EMBO J 1991 Apr;10(4):941-9

17. Brun YV, Sanfacon H, Breton R, Lapointe J. Closely spaced and divergent promoters for an aminoacyl-tRNA synthetase gene and a tRNA operon in Escherichia coli. Transcriptional and post-transcriptional regulation of gltX, valU and alaW. J Mol Biol 1990 Aug 20;214(4):845-64

18. Henkin TM. tRNA-directed transcription antitermination. Mol Microbiol 1994 Aug;13(3):381-7

19. Gourse RL, de Boer HA, Nomura M. DNA determinants of rRNA synthesis in E. coli: growth rate dependent regulation, feedback inhibition, upstream activation, antitermination. Cell 1986 Jan 17;44(1):197-205

20. Bauer BF, Kar EG, Elford RM, Holmes WM. Sequence determinants for promoter strength in the leuV operon of Escherichia coli.Gene 1988;63(1):123-34

21. Nilsson L, Vanet A, Vijgenboom E, Bosch L. The role of FIS in trans activation of stable RNA operons of E. coli. EMBO J 1990 Mar;9(3):727-34

22. Green J., M. Anjum and Guest J. R. Regulation of the ndh gene of Escherichia coli by integration host factor and a novel regulator, Arr. Microbiology. 1997; 143: 2865-2875

23. Dickerson E. R., DNA bending: the prevalence of kinkiness and the virtues of normality. Nuc. Ac. Res. 1998; 26(8):1906-1926.

24. Clariet L. and Rouviere-Yaniv J. Regulation of HU-alpha and HU-beta by CRP and FIS in *Escherichia coli*. 1996. J. Mol. Biol. 263; 126-139.

25. Mechulam Y., Blanquet S., Fayat G. Dual level control of the Escherichia coli pheST-himA operon expression: tRNA-phe-dependent attenuation and transcriptional operator-repressor control by himA and the SOS network. 1987. J. Mol. Biol. 197:453-470.

26. Miller H. I., Kirk M., Echols H. SOS induction and autoregulation of the *himA* gene for site-specific recombination in Escherichia coli. 1981. Proc. Natl. Acad. Sci. USA. 78:6754-6758.

27. Givens J. R., McGovern C., and Dombroski A. Formation of Intermediate Initiation Complex at *pfliD* and *pflgM* by sigma ^{28} RNA polymerase. 2001. J. Bac. Nov. 6244-6252.

28. Soutourina O., Kolb A., Krin E., Laurent-Winter C., Rimsky S., Danchin A. and Bertin P. Multiple Control of Flagellum Biosynthesis in Escherichia coli: Role of H-NS Protein and the Cyclic AMP Catabolite Activator Protein Complex in Transcription of the *flhDC* Master Operon. J. Bac. Dec 1999; 7500-7508.

29. Kutzukake K.Autogenous and global control of the flagellar master operon, flhD, in Salmonella typhimurium. Mol Gen Genet 1997 Apr 28;254(4):440-8.

30. Atlung T, Ingmer H. H-NS: a modulator of environmentally regulated gene expression. Mol. Microbiol 1997 Apr;24(1):7-17.

31. He B. and Zalkin H. Regulation of Escherichia coli purA by purine repressor, one component of a dual control mechanism. J Bacteriol 1994 Feb;176(4):1009-13

32. Carmona M. and Magasanik B. Activation of transcription at sigma^{54}- dependent Promoters on Linear Templates Requires Intrinsic or Induced Bending of the DNA. J. Mol. Biol. 1996. 261; 348-356.

33. Pouty M., Correa N. E. and Klose K. The novel sigma^{54} and sigma^{28}-dependent flagellar gene transcription hierarchy of Vibrio cholerae. Mol. Microbiol. (2001) 39(6), 1595-1609.

34. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 2000 Jan 1;28(1):15-8

35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997 Sep 1;25(17):3389-402

36. Jagannathan A, Constantinidou C, Penn CW. Roles of rpoN, fliA, and flgR in expression of flagella in Campylobacter jejuni. . J Bacteriol. 2001 May;183(9):2937-42.

37. Shiga Y, Sekine Y., Kano Y, and Ohtsubo E, Involvement of H-NS in transpositional recombination mediated by IS1, J Bacteriol. Apr. 2001, 2476-2484

38. Lavoie, B.D, Chaconas, G, Transposition of phague Mu DNA, Curr Top Micribiol. Immunol. (1996) 204:83-99

39. Chalmers R, Anjan Guhathakurta, Benjamin H, Kleckner N., IHF modulation of tn10 transposition: sensory transduction of supercoiling status via a proposed protein/DNA molecular spring.

40. Plaskon RR, Wartell RM. Sequence distributions associated with DNA curvature are found upstream of strong E. coli promoters. Nucleic Acids Res. 1987 Jan 26;15(2):785-96.

41. Kohno K, Wada M, Kano Y, Imamoto F., Promoters and autogenous control of the Escherichia coli hupA and hupB genes. J Mol Biol 1990 May 5;213(1):27-36

42. Balas D, Fernandez-Moreira E, De La Campa AG. Molecular characterization of the gene encoding the DNA gyrase A subunit of Streptococcus pneumoniae. J Bacteriol. 1998 Jun;180(11):2854-61.

43. Bolshoy, A. and Nevo, E. (2000) Ecologic genomics of DNA: upstream bending in prokaryotic promoters. Genome Res. 10(8):1185-1193.

44. Kalate, R.N., Kulkarni, B.D. and Nagaraja V (2002) Analysis of DNA curvature in mycobacterial promoters using theoretical models. Biophisical chem. 99, 77-97.

45. Moreno-Hagelsieb G, Collado-Vides J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics. 2002 Jul;18 Suppl 1:S329-36.

46. Satchwell SC, Drew HR, Travers AA (1986) Sequence periodicities in chicken nucleosome core DNA. J Mol Biol. 1986 Oct 20;191(4):659-75.

47. Mulligan ME, Hawley DK, Entriken R, McClure WR. (1984) Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. Nucleic Acids Res. 1984 Jan 11;12(1 Pt 2):789-800.

48. Barrios H, Valderrama B, Morett E. (1999) Compilation and analysis of sigma(54)-dependent promoter sequences. Nucleic Acids Res. 1999 Nov 15;27(22):4305-13.

49. Dlakic M, Harrington RE. (1998) DIAMOD: display and modeling of DNA bending. Bioinformatics. 1998;14(4):326-31.

50. Aviv M, Giladi H, Schreiber G, Oppenheim AB, Glaser G. (1994) Expression of the genes coding for the Escherichia coli integration host factor are controlled by growth phase, rpoS, ppGpp and by autoregulation. Mol Microbiol. 1994 Dec;14(5):1021-31.

51. Menzel R, Gellert M. (1983) Regulation of the genes for E. coli DNA gyrase: homeostatic control of DNA supercoiling. Cell. 1983 Aug;34(1):105-13.

52. Menzel R, Gellert M. (1987) Modulation of transcription by DNA supercoiling: a deletion analysis of the Escherichia coli gyrA and gyrB promoters. Proc Natl Acad Sci U S A. 1987 Jun;84(12):4185-9.

53. Froelich JM, Phuong TK, Zyskind JW. Fis binding in the dnaA operon promoter region. J Bacteriol. 1996 Oct;178(20):6006-12.

54. Polaczek P, Kwan K, Liberies DA, Campbell JL.Role of architectural elements in combinatorial regulation of initiation of DNA replication in Escherichia coli. Mol Microbiol. 1997 Oct;26(2):261-75.

55. Bahloul A, Boubrik F, Rouviere-Yaniv J. Roles of Escherichia coli histone-like protein HU in DNA replication: HU-beta suppresses the thermosensitivity of dnaA46ts. Biochimie. 2001 Feb;83(2):219-29.

56. Carmona M, Claverie-Martin F, Magasanik B. DNA bending and the initiation of transcription at sigma54-dependent bacterial promoters. Proc Natl Acad Sci U S A. 1997 Sep 2;94(18):9568-72.

57. Gralla JD. (1996) Activation and repression of E. coli promoters. Curr Opin Genet Dev. 1996 Oct;6(5):526-30.

table1

SDD	*COG*	*genes/organisms*	*Function*
Nucleotide transport and metabolism			
4.17	COG0299	25/25	phosphoribosylglycinamide formyltransferase
3.95	COG3072	8/8	adenylate cyclase
Signal transduction mechanisms			
3.38	COG1217	22/21	GTP-binding protein TypA/BipA
3.09	COG3275	10/9	Autolysin sensor kinase
Cell motility			
3.21	COG1344	38/20	flagellar hook-filament junction protein 3 FlgL
3.08	COG1360	22/19	Flagellar motor protein MotB
Transcription			
4.19	COG4977	31/18	transcriptional regulator araC family
3.81	COG0085	32/31	RNA polymerase beta subunit
3.28	COG0553	32/24	ATP-dependent RNA helicase HepA
3.16	COG1522	98/36	transcriptional regulator asnC/lrp family
Amino acid transport and metabolism			
4.66	COG4992	38/29	PLP-dependent aminotransferases
3.36	COG0253	24/24	Diaminopimelate epimerase
3.30	COG0703	25/25	shikimate kinase
3.03	COG0174	46/31	glutamine synthetase
Defense mechanisms			
3.67	COG2746	8/6	aminoglycoside N3-acetyltransferase
Cell wall/membrane/envelope biogenesis			
5.37	COG3637	29/12	outer membrane protein x precursor
5.30	COG0275	31/31	SAM-dependent methyltransferase
4.24	COG0768	52/36	penicillin-binding protein 2
4.10	COG2821	14/14	membrane-bound lytic murein transglycosylase A
3.86	COG0472	42/37	phospho-N-acetylmuramoyl-pentapeptide-transferase
3.33	COG0797	20/17	rare lipoprotein A
3.04	COG1212	15/15	3-deoxy-manno-octulosonate cytidylyltransferase
3.01	COG0770	23/23	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanyl ligase
Replication, recombination and repair			
6.81	COG0776	64/43	DNA-binding proteins HU and IHF
4.57	COG0188	45/37	DNA gyrase subunit A
4.23	COG0187	45/38	DNA gyrase subunit B
4.15	COG3385	38/6	transposase
4.08	COG1604	5/5	conserved hypothetical protein
3.13	COG0468	30/30	RecA protein
3.04	COG3611	7/7	chromosome replication initiation / membrane

			attachment protein DnaB

Translation, ribosomal structure and biogenesis

3.55	COG0268	23/23	30S ribosomal protein S20
3.50	COG0173	25/25	aspartyl-tRNA synthetase
3.28	COG0012	29/29	GTP-binding protein
3.22	COG4108	18/18	peptide chain release factor 3
3.04	COG5256	7/7	translation elongation factor EF-1, subunit alpha (tuf)

Posttranslational modification, protein turnover, chaperones

3.27	COG0719	37/24	Iron-regulated ABC-type transporter membrane component (SufB)

Inorganic ion transport and metabolism

4.19	COG1392	19/18	conserved hypothetical protein
3.47	COG1553	12/9	putative ACR involved in intracellular sulfur reduction

Cell cycle control, cell division, chromosome partitioning

4.9	COG3116	11/11	cell division protein (FtsL)
3.56	COG0849	21/21	cell division protein (ftsA)
3.48	COG0552	29/29	cell division protein (ftsY)
3.20	COG3096	5/5	cell division protein (mukB)
3.11	COG3006	5/5	killing factor protein (KICB)
3.02	COG3095	5/5	killing protein supressor (kicA, mukE)

Carbohydrate transport and metabolism

3.36	COG0205	25/23	6-phosphofructokinase
3.01	COG0166	26/26	glucose-6-phosphate isomerase

General function prediction only

3.82	COG4572	5/5	cation transport regulator ChaB
3.82	COG1084	8/8	GTP-binding protein, GTP1/OBG-family
3.70	COG1075	14/12	triacylglycerol lipase precursor
3.33	COG0795	27/14	putative membrane protein
3.18	COG2071	16/16	glutamine amidotransferase, class I
3.12	COG3081	7/7	Nucleoid-associated protein
3.03	COG2607	9/9	conserved hypothetical protein
3.02	COG1823	8/8	sodium-glutamate symporter

Function unknown

5.6	COG2001	23/23	conserved hypothetical protein
4.18	COG3870	7/7	hypothetical nitrogen regulatory protein P-II (GLNB)
4.00	COG3862	5/5	predected metal-binding protein
3.85	COG1799	14/14	hypothetical protein
3.73	COG3025	10/10	conserved hypothetical protein
3.70	COG1945	9/8	hypothetical protein
3.65	COG2302	9/9	conserved hypothetical protein
3.47	COG0779	21/21	conserved hypothetical protein
3.41	COG4095	6/5	conserved hypothetical protein
3.32	COG2976	10/10	hypothetical protein

3.24	COG0762	17/17	conserved hypothetical protein
3.15	COG4807	7/7	conserved hypothetical protein
3.08	COG3665	6/5	hypothetical protein
3.01	COG4649	5/5	conserved hypothetical protein

Table2

SDD	***COG***	***genes/organisms***	***Function***
Replication, recombination and repair			
2.84	COG4644	9/6	Tn5044 transposase
2.76	COG1943	45/13	transposase
2.56	COG1315	5/5	hypothetical protein
2.41	COG3464	30/6	IS1167, transposase
2.37	COG0582	113/51	integrase/recombinase
2.29	COG3666	23/8	transposase-IS1562
Translation, ribosomal structure and biogenesis			
2.76	COG0124	27/26	histidyl-tRNA synthetase
2.75	COG0806	21/21	16S rRNA processing protein
2.24	COG0689	18/18	ribonuclease PH
2.10	COG0228	20/20	30S ribosomal protein S16
2.03	COG1544	19/19	Ribosome-associated protein Y
2.03	COG0060	24/23	isoleucyl-tRNA synthetase
2.01	COG0023	9/9	Translation initiation factor (SUI1 related)
2.00	COG0198	23/23	50S ribosomal protein L24
1.97	COG0081	24/24	50S ribosomal protein L1
Posttranslational modification, protein turnover, chaperones			
2.83	COG0326	17/17	heat shock protein (HtpG)
2.25	COG0484	25/24	DnaJ protein
Cell motility			
2.35	COG4787	7/7	flagellar basal-body rod protein (flgF)
2.30	COG1291	16/15	Flagellar motor component (MotA)
2.26	COG1377	15/14	polar flagellar assembly protein (FlhB)
2.25	COG1558	13/13	flagellar basal-body rod protein (FlgC)
2.00	COG1261	10/10	flagella basal body P-ring formation protein
Transcription			
2.84	COG2901	9/9	DNA-binding protein (Fis)
2.46	COG1758	22/22	RNA polymerase omega subunit
2.18	COG2002	16/12	transcription regulator (spoVT)
2.13	COG1508	15/14	RNA polymerase sigma-54 factor
2.05	COG0202	23/23	RNA polymerase alpha subunit
Amino acid transport and metabolism			
2.16	COG0159	21/21	tryptophan synthase alpha chain
2.08	COG0133	20/20	tryptophan synthase beta chain

Cell wall/membrane/envelope biogenesis

2.24	COG0357	19/19	glucose inhibited division protein B
2.11	COG3951	7/7	putative flagellar protein
2.11	COG1589	16/16	cell division protein (ftsQ)

Cell cycle control, cell division, chromosome partitioning

2.87	COG0772	40/31	cell division protein (ftsw)
2.77	COG3115	7/7	cell division protein (ZipA)
2.54	COG4839	5/5	cell division protein (FtsL)
2.17	COG0445	19/19	glucose inhibited division protein A

Fig. 1

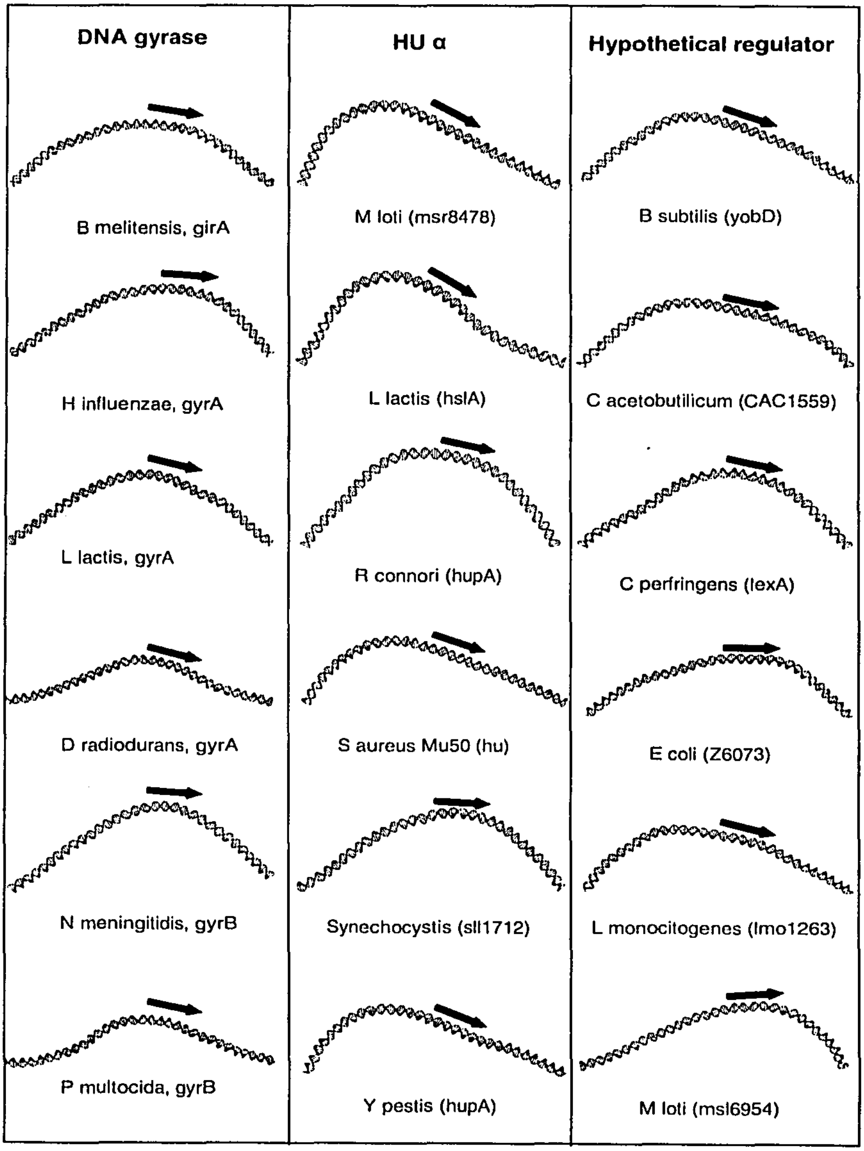
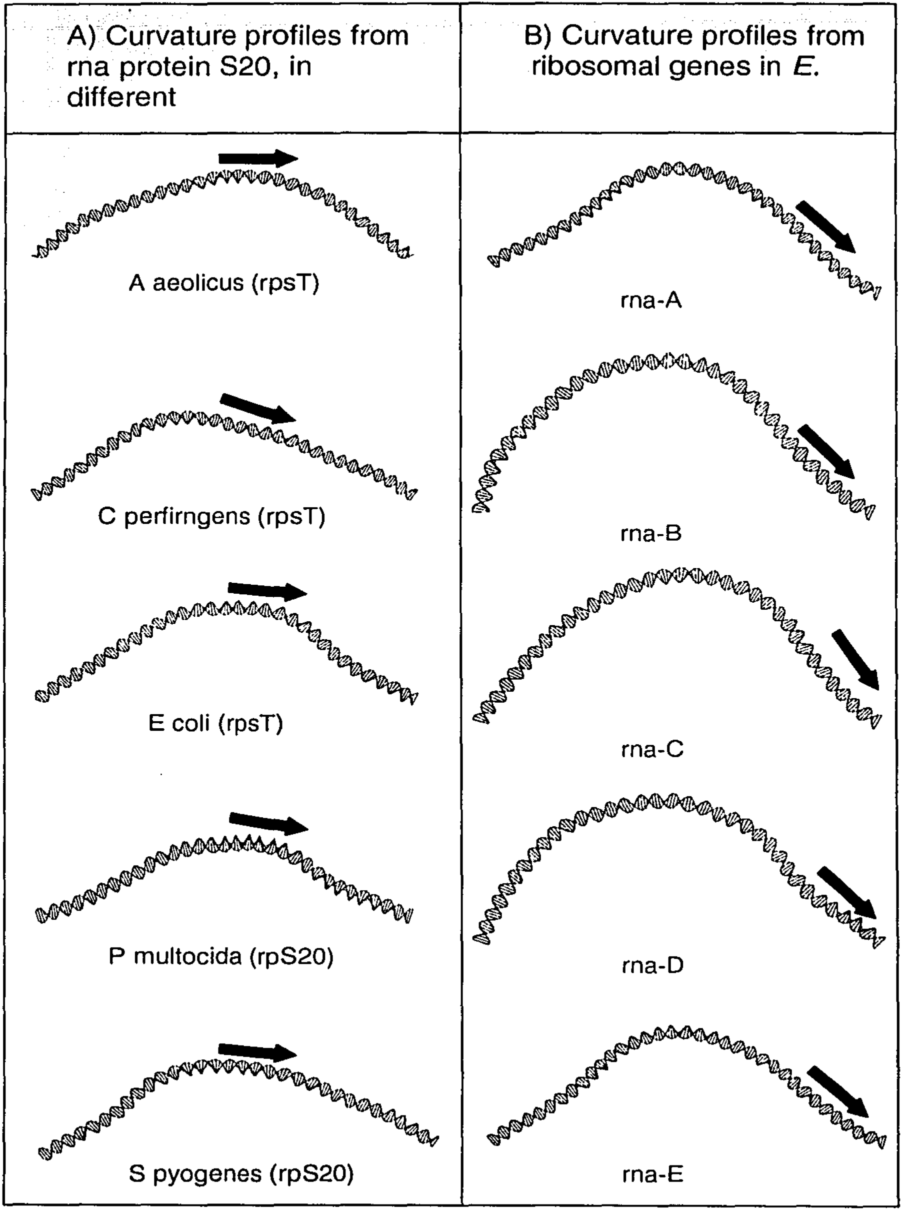

DNA gyrase
B melitensis, girA
H influenzae, gyrA
L lactis, gyrA
D radiodurans, gyrA
N meningitidis, gyrB
P multocida, gyrB
HU α
M loti (msr8478)
L lactis (hslA)
R connori (hupA)
S aureus Mu50 (hu)
Synechocystis (sll1712)
Y pestis (hupA)
Hypothetical regulator
B subtilis (yobD)
C acetobutilicum (CAC1559)
C perfringens (lexA)
E coli (Z6073)
L monocitogenes (lmo1263)
M loti (msl6954)

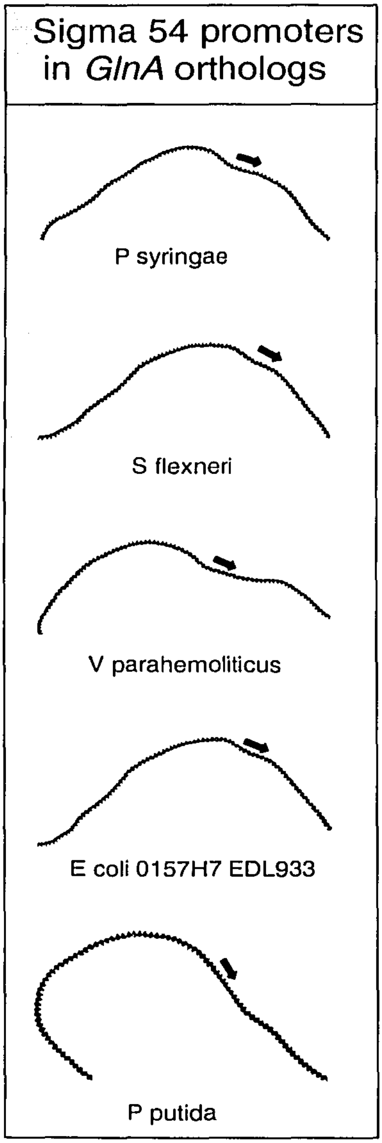
Fig. 2



A) Curvature profiles from rna protein S20, in different

B) Curvature profiles from ribosomal genes in *E.*

Fig. 3


Sigma 54 promoters in *GlnA* orthologs
P syringae
S flexneri
V parahemoliticus
E coli 0157H7 EDL933
P putida

Conclusiones.

Determinantes globales.

Nuestros principales resultados sobre el análisis de determinantes globales de la curvatura del DNA genómico demostraron que:

- Cada organismo presenta un perfil de curvatura único y no aleatorio. La permutación de los nucleótidos de un genoma produce un cambio significativo en el perfil de curvatura del mismo.
- La curvatura de un genoma depende parcialmente de su contenido de GC. Aunque se ha observado que el porcentaje de GC en una secuencia dada imprime un bias en el perfil de curvatura, genomas naturales con el mismo contenido de GC presentan perfiles de curvatura muy diferentes.
- Las regiones intergénicas son, en promedio, más curvas que las regiones codificantes.
- Existen genomas hipercurvos, cuyo perfil es más curvo que lo esperado al azar, y genomas hipocurvos, cuyo caso es el contrario.
- El uso de codones se halla íntimamente relacionado con la curvatura del genoma; organismos con genomas hipercurvos utilizan preferencialmente codones que favorecen valores altos de curvatura.
- La fracción de aminoácidos del proteoma también se haya relacionada con la curvatura; organismos con genomas hipercurvos utilizan prioritariamente aminoácidos cuyos codones favorecen la ocurrencia de DNA curvo.

Señales conservadas en regiones de regulación.

Dentro del análisis de la curvatura del DNA en relación con la regulación transcripcional, las principales conclusiones obtenidas son:

- Existen grupos de genes ortólogos en diferentes genomas que presentan señales de curvatura estadísticamente significativas en sus regiones reguladoras, indicando un posible mecanismo común de regulación.
- Varios casos de genes regulados por curvatura, experimentalmente determinados en un solo organismo, presentan mecanismos similares de regulación en una gran variedad de genomas.
- Hemos identificado nuevos grupos de genes para los cuales no hay evidencia experimental previa, que también presentan señales de curvatura conservadas. En estos casos nuestras predicciones pueden guiar el estudio experimental de los mecanismos de regulación transcripcional.

El estudio de la estructura de la molécula de DNA en las regiones de regulación es fundamental para comprender los mecanismos de regulación, que en muchos casos poseen información independiente de la secuencia.

Este trabajo ha culminado en la elaboración de un manuscrito a ser sometido para su publicación en una revista de arbitraje internacional (Nucleic Acids Research).

Resultados adicionales.

Actualmente estamos desarrollando una metodología experimental para verificar si la curvatura del DNA juega un papel en la regulación de genes que presentan señales significativas de curvatura y cuya expresión depende de la misma proteína reguladora, tal es el caso de los genes dependientes del regulador TyrR. No existen en la literatura datos que relacionen el mecanismo de acción de TyrR con la presencia de regiones curvas en el DNA.

Estudios previos han propuesto que el inicio de la transcripción de los genes dependientes del factor sigma 54 depende a su vez de la proteína IHF o bien de un sitio curvo en el DNA para permitir el contacto de la región "enhancer" con el promotor (12). Con el fin de verificar si existe un motivo de curvatura conservado en las regiones promotoras de estos genes, examinamos los perfiles de curvatura de los promotoras caracterizados experimentalmente, identificando regiones curvas en los genes relacionados con la biosíntesis de la glutamina *glnA y glnB* en *B. japonicum, S. typhimurium* y *E. coli.* La presencia de señales de curvatura en la región promotora de *glnA* está descrita experimentalmente (27) y se analiza en el manuscrito "Conservación de señales de curvatura en regiones de regulación" incluido en el desarrollo de esta tesis, en donde se demuestra la conservación de estas señales en 31 organismos.

Perspectivas.

La capacidad de identificar estructuras conservadas de forma independiente de la secuencia nos abre varias posibilidades de estudio muy prometedoras. Una primera propuesta de continuación de este proyecto podría involucrar la identificación de señales de curvatura conservadas dentro de un solo genoma, y relacionarlas con reguladores específicos de procesos celulares. Los primeros estudios en esta dirección nos han permitido identificar sitios de integración de transposones en el genoma de *E. coli*.

Otra posible línea de investigación podría adaptar las metodologías desarrolladas para investigar señales conservadas en genomas virales. Los genomas virales poseen características geométricas muy particulares debido a su tamaño pequeño, y se han identificado señales de DNA curvo relacionadas con varios procesos del metabolismo viral (19). Hasta ahora no se ha realizado ningún estudio global en este sentido, a pesar de que se hallan disponibles más de 700 genomas virales completos en el GenBank.

Una línea de interés, que también tiene algunos avances en casos aislados, implica la caracterización de las estructuras de DNA de promotores de genes eucariontes. Reportes previos han identificado motivos de regulación curvos en algunos genes en *H. sapiens* (20, 21). El estudio de la curvatura podría aportar avances significativos en la comprensión de los mecanismos de regulación transcripcional y de splicing en eucariontes.

Otra aplicación inmediata de estos datos podría ser la construcción de una base de datos que integre la información obtenida a través de la estructura del DNA en relación con rutas metabólicas o datos de regulación transcripcional.

Una vertiente más de este trabajo sería caracterizar los sitios de alta flexibilidad en el DNA, estudiar su distribución en el genoma y su relación con los procesos biológicos. Actualmente hay algunos trabajos pioneros en la caracterización de la flexibilidad del DNA, la mayoría en casos aislados (22). Este proyecto completaría una descripción muy general sobre los parámetros estructurales del DNA relevantes en la Biología.

Bibliografía.

1. Harrington, R. E., (1992) Mol. Microbiol. 6. 2549-2555.
2. Marini J. C., Effron P. N., Goodman T. C., Singleton C. K., Wells R. D., Wartell R. M., Englund P. T. J (1984) Biol. Chem. 259(14). 8974-8979.
3. De Santis, P., Palleschi, A., Savino, M., Scipioni, A. (1990), Biochemistry. 29, 9269-9273.
4. Calladine, C.R., Drew, H. R., McCall, M. J., (1988) J. Mol. Biol. 201, 127-137.
5. Bolshoy, A., McNamara, P., Harrington, R. E., Trifonov, E. N., (1991) PNAS, 88,2312-2316.
6. Satchwell, S.C., Drew, H. R., Travers, A. A., (1986) J. Mol. Biol. 191, 659-675.
7. Hagerman, P. J., (1990) Annu. Rev. Biochem. 59. 755-781.
8. Gabrielian A., Vlahoviеck K., Pongor S. (1997) FEBS Lett. 406(2). 69-74.
9. Goodsell, D. S., Dickerson, R. E., (1994) Nuc. Ac. Res. 22(24). 5497-5503.
10. Jauregui R., O'Reilly F., Bolivar F., Merino E. (1998) Microb. Comp. Genomics. 3(4). 243-53.
11. Jauregui R., Bolivar F., Merino E. (2000) Microb. Comp. Genomics. 5(1). 7-15.
12. Carmona, M., Magasanik, B. (1996) J. Mol. Biol. 261. 348-356.
13. Espinosa-Urgel, M., Tormo, A. (1993) Nuc. Ac. Res. 21(16). 3667-3670.
14. Pérez-Martín, J., Rojo, F., De Lorenzo, V. Microbiol. Rev. (1994) 58(2). 268-290.
15. Wojtuszewski, K., Hawkings, M. E., Cole, J. L., Mukerji, I. (2001) Biochemistry. 40. 2588-2598.
16. Gabrielian, A., Landsman, D., Bolshoi, In Silico Biol. (2000).1(4).183-196.
17. Bolshoy. A., Nevo, E. (2000) Genome Res. 10(8). 1185-93.
18. Kalate, R. N., Kulkarni, B. D., Nagaraja, V., (2002), Biophis. Chem. 99. 77-97.
19. Bask, S., Olsen, L., Hattman, S., nagaraja, V. (2001) J. Bio. Chem. 276 (23). 19836-198344.
20. Martins R. P., Ujfalusi A. A., Csiszar K., Krawetz S. A. DNA Seq. 12(4). 215-27.
21. Li X. M., Onishi Y., Kuwabara K., Rho J. Y., Wada-Kiyama Y., Sakuma Y., Kiyama R. (2002) Gene. 294(1-2). 279-90
22. Pedersen, A. G., Jensen L. J., Brunak, S., Staerfelt, H. H., Ussery, D. W. (2000) J. Mol. Biol. 299. 907-930.
23. Atlung T, Ingmer H. H-NS: a modulator of environmentally regulated gene expression. Mol. Microbiol 1997 Apr;24(1):7-17.
24. Trifonov EN. Curved DNA. CRC Crit Rev Biochem 1985;19(2):89-106
25. Collis CM, Molloy PL, Both GW, Drew HR. Influence of the sequence-dependent flexure of DNA on transcription in E. coli. Nucleic Acids Res 1989 Nov 25;17(22):9447-68
26. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000 Jan 1;28(1):33-6
27. Carmona M, Claviere-Martin F and Magasanik B. DNA bending and the initiation at sigma 54 – dependent bacterial promoters. Proc. Nat. Acad. Sci. USA. (1997) 94, 9568-9572.