Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA

ALLAN CAMPBELL†‡, JAN MRÁZEK§, AND SAMUEL KARLIN§

Departments of † Biological Sciences and § Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Allan Campbell, May 26, 1999

ABSTRACT Our basic observation is that each genome has a characteristic "signature" defined as the ratios between the observed dinucleotide frequencies and the frequencies expected if neighbors were chosen at random (dinucleotide relative abundances). The remarkable fact is that the signature is relatively constant throughout the genome; i.e., the patterns and levels of dinucleotide relative abundances of every 50-kb segment of the genome are about the same. Comparison of the signatures of different genomes provides a measure of similarity that looks at all the DNA of an organism and does not depend on the ability to align homologous sequences of specific genes. Genome signature comparisons show that plasmids, both specialized and broad-range, and their hosts have substantially compatible (similar) genome signatures. Mammalian mitochondrial (Mt) genomes are very similar, and animal and fungal Mt are generally moderately similar, but they diverge significantly from plant and protist Mt sets. Moreover, Mt genome signature differences between species parallel the corresponding nuclear genome signature differences, despite large differences between Mt and host nuclear signatures. In signature terms, we find that the archaea are not a coherent clade. For example, Sulfolobus and Halobacterium are extremely divergent. There is no consistent pattern of signature differences among thermophiles. More generally, grouping prokaryotes by environmental criteria (e.g., habitat propensities, osmolarity tolerance, chemical conditions) reveals no correlations in genome signature.

Extensive data support the proposal that each living organism possesses a genomic signature consisting of dinucleotide relative abundance values calculated from genomic sequences (1–3). Explicitly, the genomic signature profile consists of the array \( \{p_{XY}\} \), where \( p_{XY} = f_{XY}/f_X f_Y \), where \( f_{XY} \) denotes the frequency of the mononucleotide \( X \) and \( f_{XY} \) the frequency of the dinucleotide \( XY \), both computed from the sequence concatenated with its inverted complement. These dinucleotide relative abundance values \( \{p_{XY}\} \) minus 1 (also termed dinucleotide biases) effectively assess differences between the observed dinucleotide frequencies and those expected from random associations of the component mononucleotide frequencies. From data simulations and statistical theory, the estimates \( \hat{p}_{XY} \approx 0.78 \) or \( \hat{p}_{XY} \approx 1.23 \) convey significant underrepresentation or overrepresentation, respectively, for 50-kb random DNA contigs (1–3). We present substantial data showing that the genome signatures of bacterial plasmids are pervasively similar to those of their natural hosts. By contrast, the signatures of animal mitochondrial DNA are not close to those of their hosts but are generally concordant with those of other animal mitochondrial (Mt) DNA.

Justifications for Using Genome Signature. Biochemical experiments in the 1960s and 1970s measuring nearest-neighbor frequencies (4, 5) established that the set of dinucleotide relative abundance values \( \{p_{XY}\} \) is a remarkably stable property of the DNA of an organism. From this perspective, the set of dinucleotide relative abundance values constitutes a genomic signature that is diagnostic and can discriminate sequences from different organisms (3, 6). What causes the uniformity of signature throughout the genome? It pervades both noncoding and coding DNA (7), and hence cannot be explained by preferential codon usage. A reasonable explanation postulates differences in the replication and repair machinery of different species, which either preferentially generate or preferentially select specific dinucleotides in the DNA. These effects might operate through local DNA structures (base step conformational tendencies), context-dependent mutation rates, methylation, and/or other DNA modifications (1–3, 6).

A measure of genomic signature difference between two sequences \( f \) and \( g \) (from different organisms or from different regions of the same genome) is the average absolute dinucleotide relative abundance difference calculated as

\[
\delta^* (f, g) = 1/16 \sum |p_{XY}^*(f) - p_{XY}^*(g)| \quad (\delta^* \text{ difference})
\]

where the sum extends over all dinucleotides. Levels of \( \delta^* \) differences of 50-kb contigs for some reference examples are described in the Introduction of the accompanying paper (8).

Genome Signature Comparisons. Fig. 1 displays the genome signature profiles for all currently available extensive genome sequence sets. These include 22 eubacterial genome ensembles (mostly complete genomes), 5 archael chromosomes, the Saccharomyces cerevisiae and Caenorhabditis elegans complete genomes, and extensive nonredundant sequence collections from 5 other eukaryotic species. For each genome, the dinucleotide biases \( \{p_{XY}\} \) are substantially invariant (9).

The dinucleotide TA is broadly underrepresented or low normal in prokaryotic sequences about the level \( 0.50 \leq p_{TA}^* \leq 0.82 \) (exceptions include Rickettsia prowazekii, Clostridium acetobutylicum, and the archaea P. aerophilum, P. horikoshii, and Sulfolobus sp. (full names appear in the legend to Figs. 1 and 2). TA underrepresentation is also pervasive in eukaryotic chromosomes (exception P. falciiparam) but not in eukaryotic small viral genomes or in animal mitochondrial genomes (10, 11).

Among prokaryotes, CG is suppressed (underrepresented) in M. genitalium (but not in M. pneumoniae), in R. prowazekii, in B. burgdorferi, in C. jejuni, in the low-G+C Gram-positive sequences of Streptococcus and Clostridium, and in several thermophiles, including M. jannaschii, M. thermoautotrophium, Sulfolobus spp., but not in P. aerophilum or P. horikoshii. At the other extreme, CG is overrepresented in B. steatorthermophilus, in halobacteria, and also in several \( \beta \)- and \( \alpha \)-proteobacterial genomes (e.g., Neisseria spp. and Rhizobium spp.). Among eukaryotes, CG shows potent suppression in vertebrates (even in deuterostomes). Overall, \( p_{CG} \) values in verte-
brates range from 0.23 to 0.40, whereas they are in the normal range for insects, worms, and most fungi. It has been shown (12) that unmethylated CG dinucleotides of normal frequency in most enteroproteobacteria can induce an immune response in mammalian genomes, where CG is very low. Is this a concomitant of genomic signature biases? The reverse dinucleotide, GC, is predominantly overrepresented in many \( \beta \)- and \( \gamma \)-proteobacterial sequences and in several low-G+\( \chi \) Gram-positive bacterial genomes (e.g., \( B. \) \( \text{subtilis} \) and \( C. \) \( \text{acetobutylicum} \); Fig. 1). In eukaryotes, moderate overrepresentation of GC occurs in \( Drosophila \) species but apparently not in other metazoa.

**TT/AA** is overrepresented in several proteobacteria, in Mycoplasmas, in \( \text{Synechocystis} \) sp., in \( \text{Deinococcus radiodurans} \), in \( A. \) \( \text{aeolicus} \) among prokaryotes, and in insects and worms among eukaryotes. There are no underrepresentations of TT/AA. Overrepresentations of GC occurs in \( Drosophila \) species but apparently not in other metazoa.

**d**\(^{\ast} \) Differences Among Prokaryotic Sequences. Fig. 2 reports average \( d^{\ast} \) differences based on multiple disjoint 50-kb contigs among prokaryotic genomic sets, each at least 100 kb in aggregate length and most genome sequence collections exceeding 800 kb. The \( d^{\ast} \) differences for all pairs of contigs for every pair of genomes are highly stable with limited variation (cf. refs. 1–6).

(i) Within-species \( d^{\ast} \) differences (diagonal elements of Fig. 2) range from 12 to 52 (all \( d^{\ast} \) differences are multiplied by 1000), and average between species \( d^{\ast} \) differences range from 34 to 340.

(ii) The rickettsial sequences are grouped with \( \alpha \)-proteobacteria, apparently on the basis of rRNA gene comparisons. Is this consistent? The classical \( \alpha \)-proteobacterial types are divided into two major subgroups: \( \alpha_{1} \), including \( R. \) \( \text{rubrum} \) spp., that function importantly in nitrogen fixation, and \( \alpha_{2} \), including \( R. \) \( \text{denitrificans} \), found predominantly in soil and marine habitats and doing anoxygenic photosynthesis. A tentative third group, \( \alpha_{3} \), includes the \( R. \) \( \text{rickettsii} \) and \( E. \) \( \text{ehrlichia} \) clades (obligate intracellular parasites). Genome signature comparisons indicate drastic discrepancies between the combined groups \{\( \alpha_{1}, \alpha_{2} \)\} and the group \( \alpha_{3} \). Moreover, the \( \alpha_{1} \) and \( \alpha_{2} \) genomes are pervasively of high G+\( \chi \) content (\( \geq \)60%), whereas \( \alpha_{3} \) genomes are of low G+\( \chi \) content (\( \leq \)32%). The \( d^{\ast} \) differences between \( R. \) \( \text{rickettsii} \) and classical \( \alpha \)-proteobacterial sequences are generally \( >200 \), very distant (see Fig. 2).

(iii) \( \text{Sulfolobus} \) spp. \( d^{\ast} \) differences from other prokaryotes. Fig. 2 includes seven archaeal sequence collections consisting of one halobacterial genus (HALSPA) (see legends to Figs. 1 and 2).
and 2 for full names, and we use the Swiss-Prot abbreviations),
two methanogens (METJA, METTH), one “derived” methanogen (ARCFU),
and three archaeal thermophiles (PYRHO, PYRAE, SULSP). Thermophiles (exception,
PYRHO) tend to be relatively closer to vertebrate eukaryotes
than to eubacterial sequences, with
d* differences in the range
100–150 (3, 9), whereas HALSP have generally
d*.

Thus, the highly diverse
d* differences indicate that a consis-
tent description for the archaeal sequences is problematical.

We summarize the genomic signature contrasts for
Sulfolobus
in the following array.

\[ d^*(\text{Sulfolobus}, \text{Clostridium}) \approx 85, \text{moderately similar};
d^*(\text{Sulfolobus}, \text{Rickettsia}, \text{and Buchnera}) \approx 125–130, \text{distantly similar};
d^*(\text{Sulfolobus}, \text{other thermophilic archaea}) \approx 71–114,\]

weakly to \text{distantly similar};
d^*(\text{Sulfolobus}, \text{purple proteobacteria and high } G+C \text{ Gram-positive}) \approx 190–270, \text{very distant};
d^*(\text{Sulfolobus}, \text{cyanobacteria}) \approx 145–177, \text{distant}.

(iv) Halobacterial genome sequences are outliers. Intrigu-
ingly, with respect to genome signature comparisons, the
closest to Halobacterium spp. are the Streptomyces sequences,
weakly similar. The \text{d* differences of Halobacterium} spp. from
other prokaryotes modestly exceed the extreme level of 250.

**Genome Compatibility Between Plasmids and Host.** Plas-
mids are genetic mobile elements among bacterial cells, gen-
erally laterally transferred by conjugation (on occasion by
transduction or transformation). Plasmids carry restriction
systems, antibiotic resistance genes, heavy-metal cofactors, Nif
(nitrogen fixation) genes, and other contingency functions.
Replication of plasmids is largely governed by host machinery.
How the genome signature of a plasmid sequence compares
with that of its host sequence is assessed for available pro-
karyotic genomes with completely sequenced plasmids. Fig. 3
reports average \text{d* differences for each plasmid compared with}
representatives of 52 prokaryotic genomic collections.

Importantly, the \text{d* differences between each plasmid and its natural
host rank among the closest. Thus, the examples of Fig. 3 are
consistent with the proposition that plasmids are viable in
bacterial hosts only when their genome signatures are suffi-
ciently compatible (moderately similar, \(55 \leq \text{d*} \leq 85\), or at least
weakly similar, \(90 < \text{d*} < 115\)) with the host genome signature
or they can rapidly be made similar.

**B. burgdorferi.** The complete genome extends 0.911 Mb and
contains 17 plasmids, of which 11 have been sequenced from
10 kb to 54 kb in size labeled A to K. The average
\text{d* difference within the} B. burgdorferi (BORBU) genome is 25 with range
3–66. The \text{d* differences among all plasmids and compared
with a complete set of 50-kb contigs of the host genome range
from 42 to 84, clearly moderately similar (data not shown).

Mutual plasmid sequence \text{d* differences show
moderate to weak similarity.}

Specific plasmids. M. jannaschii, mja. This archaeal genome
contains two plasmids of 58 kb (called large) and 16.5 kb
(smaller). Again, we find that the two plasmids are mutually close
(\text{d*} = 3\text{5}) and each moderately similar to the host genome. Rho-
Zobium spp., rhi. We have available about 150 kb of
nonredundant aggregate sequences from R. leguminosarum
and about 250 kb of sequence from R. meliloti. There is a
mammoth plasmid of about 550 kb. Partitioning these se-
quencies into 50-kb contigs yields the average \text{d* differences given in Fig. 3 of close similarity. Halobacterium} spp., hal. An

---
Broad-host-range plasmids. Two plasmids that stably replicate in many hosts are RSF1010 and RP4 (13, 14); see legend to Fig. 3. We list next those host prokaryotic species with at least 50 kb of available genome sequence which accept stably these plasmids. RSF1010: *Escherichia coli*, *Pseudomonas aeruginosa*, *Pseudomonas putida*, *Azotobacter vinelandii*, *Rhizobium meliloti*, *Agrobacterium tumefaciens*, *Escherichia coli*, *Pseudomonas putida*, *Rhodospirillum rubrum*, *Shigella boydii*, *Salmonella typhimurium*, *Vibrio cholerae*, and *Rhizobium meliloti*.

The broad-range plasmids are generally moderately or weakly similar to the other plasmids from proteobacterial genomes. The plasmid from *L. lactis* is distinctly similar to the broad-range plasmids. The plasmids from *B. burgdorferi*, *Halobacterium* spp., and *M. jannaschii* are very distant from all proteobacterial plasmids. The δ* differences among *B. burgdorferi* plasmids are mutually closely or moderately similar, except plasmid G, which is only weakly similar. Plasmid G is close (δ* = 49) to the *L. lactis* plasmid, perhaps implying that a low G+C Gram-positive bacterium may be the source of plasmid G. The two broad-range plasmids are mutually moderately similar (δ* = 83). This similarity might mean either that relatively close genome signatures promote plasmid establishment or that the plasmids have acquired their hosts' signatures during long-term residence. Experiments on conjugation can address issues such as specificity vs. wide host range and relevance of size and signature for plasmid compatibility. We interpret the similarities in signature between plasmids and their bacterial hosts as implying that they share much replication and repair machinery, perhaps because the prokaryotic cell is not compartmentalized to the degree that the eukaryotic cell is.

**Genomic Signature δ* Differences Among Mitochondrial Mt Genomes.** Fig. 4 details δ* differences among a broad spectrum of Mt genomes. A succinct summary is given in Table 1. Mammalian Mt genome signatures appear almost as random samples of each other. Why are the genome signatures among vertebrate Mt sequences significantly more similar than those among the corresponding host nuclear genomic sequences (15)? δ* differences among organisms putatively reflect variations in replication and/or repair (1–3, 6). The replication machinery for animal Mt DNA apparently varies less than that for host DNA, perhaps because Mt replication is less affected by changes in external environment or developmental programs than is host DNA replication. δ* differences of Mt from protostomes (e.g., insect, mollusk) versus deuterostomes show δ* differences mostly in the range 90–115, weakly similar. The worm (LUMTE and *C. elegans*) genomes compared with deuterostome Mt are closely to weakly similar, δ* = 41–101. The fungal Mt sets (excluding *S. cerevisiae*) compared with deuterostome Mt also yield δ* differences in the range 43–114. The *S. cerevisiae* mtDNA (~78 kb) composition is an extreme anomaly attested to by the inordinate δ* differences from all other Mt or nuclear genomes, contributed to by about 100

![Image](image-url)
G+C-rich clusters, each about 50 to 100 bp long, separated by A+T-rich spacers and numerous transposable elements. The two plant Mt genomes (ARATH, MARPO, see Fig. 4), of about 150-kb length, compared with animal Mt are weakly similar to distant. The green alga CHLRE Mt is very distant from most other Mt sequences (d ∗ > 180). The other two algal Mt sequences (PROWI, CHOCR) are generally weakly similar to nonprotist Mt sequences. The protist Mt divide into two subgroups, P1 = (RECAM, ACACA) and P2 = (PARAU, TRYBR, LEITA). The RECAM Mt genome signature compared with deuterostome Mt is weakly to distantly similar. ACACA is a bit farther (d ∗ = 94–144). The protist group P2 are mostly distant to very distant from animal Mt sequences. The plant A. thaliana Mt sequence is distant from mammalian Mt sequences, possibly suggesting a polyphyletic ancestry separating metazona, plant, and protist Mt.

Animal Mt sequences show significant underrepresentations of CG dinucleotides, pCG ≈ 0.40–0.60 (10), about to the same extent as occurs in vertebrate nuclear genomic sequences. Virtually all animal Mt maintain normal representations of TA dinucleotides, whereas the corresponding nuclear DNAs predominantly have TA in low relative abundance, suggesting that mtDNA may be thermodynamically less stable than nuclear DNA because the dinucleotide TA has the lowest stacking energies compared with all other base steps. The fungal Mt of S. pombe has pCG ≈ 0.54, typical of animal Mt. However, the Podospora anserina fungal Mt genome has pCG in the normal range. The Mt genome of A. thaliana has pCG = 0.73, significantly low. The single persistently high p value occurs for pCC/GG ≈ 1.30 in animal and fungal Mt sequences. d ∗ differences between species parallel the corresponding nuclear d ∗ differences, despite large differences between Mt and corresponding host nuclear signatures (15).

**Discussion**

Prokaryotic molecular taxonomy heretofore has been derived predominantly from sequence comparisons among rRNA genes. There are many uncertainties and controversies regarding divisions among prokaryotes (for recent reviews, see refs. 16 and 17). Protein sequence comparisons and associated phylogenetic tree constructions are even more conflicting relative to evolutionary relationships (18, 19).

**Table 1. Summary of d ∗ differences of Mt genomes**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>d ∗</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among mammalian Mt</td>
<td>&lt;40</td>
<td>Very close</td>
</tr>
<tr>
<td>Among all vertebrates including birds, frogs, fish</td>
<td>&lt;67</td>
<td>Mainly close</td>
</tr>
<tr>
<td>Vertebrates vs. nonvertebrate deuterostomes</td>
<td>55–120</td>
<td>Moderate to weak similarity</td>
</tr>
<tr>
<td>Deuterostomes vs. protostomes</td>
<td>Mostly 34–152</td>
<td>Moderate to distant similarity</td>
</tr>
<tr>
<td>Fungi vs. animals</td>
<td>Mostly 43–128</td>
<td>Moderate to weak similarity</td>
</tr>
<tr>
<td>Plants vs. (animals and fungi)</td>
<td>84–179</td>
<td>Weak similarity to distant</td>
</tr>
<tr>
<td>C. reinhardtii vs. animals</td>
<td>Mostly 141–212</td>
<td>Distant to very distant</td>
</tr>
</tbody>
</table>

**Fig. 4.** Average d ∗ differences among complete Mt genomes based on a single sequence sample for genomes of <60 kb and multiple samples of about 50 kb for larger genomes.
Conventional methods of phylogenetic reconstruction from sequence information employ only similarity or dissimilarity assessments of aligned homologous genes or regions. Some difficulties intrinsic to this approach compared to the use of genome signature include the following: (i) Alignments of distantly related long sequences (e.g., complete genomes) are generally not feasible for various reasons, including chromosomal rearrangements, whereas signature comparisons do not depend on alignments. (ii) Different phylogenetic reconstructions (trees) may result for the same set of organisms based on analysis of different protein, gene, or noncoding sequences. Attempts to overcome these conflicts by “averaging” over many proteins are problematic because of biases in species sampling, effects of lateral transfer, complications of gene duplications, and inadequacies and artifacts of phylogenetic methods. As the signature has remarkably low variance throughout the genome, a tree based on δe differences is independent of which genome segments of 50 kb (or longer) is used in its construction. (iii) Chimeric origins and lateral transfer between distantly related organisms complicate alignment-based phylogenies. Signature comparisons are unaffected by these factors. On the other hand, alignment-based comparisons may provide information on individual gene origins that the signature method does not. (iv) Tree construction derived from aligned sequences cannot be applied to organisms for which similar gene sequences are largely unavailable (e.g., for bacteriophages, diverse eukaryotic viruses). (v) That “lateral transfer” is pervasive among prokaryotic genomes is now widely appreciated. Vectors for lateral transfer are exogenous transposons, hitchhiking on plasmids or chromosomes, whereas signature comparisons do not ameliorate even more rapidly, for both plasmids and laterally transferred chromosomal segments. Whereas most successful gene transfer between lineages is very likely intraspecific, an appreciable amount of transfer among distantly related bacterial species seems to have accumulated over time (22, 23). Despite such transfer, the signatures currently observed are almost invariant throughout the genome. Some means based on codon usage biases for ascertentions of laterally transferred genes in bacterial organisms are set forth in ref. 23.

We are happy to acknowledge valuable discussions and comments on the manuscript by Drs. B. Edwin Blaisdell, Dale Kaiser, and David Relman. S.K. is supported in part by National Institutes of Health Grants 5R01GM10452-34 and 5R01HG00335-11 and National Science Foundation Grant DMS9704552.