Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 12832–12836, December 1994
Evolution

Comparisons of eukaryotic genomic sequences

(dinucleotide relative abundance/molecular evolution/stacking energies)

SAMUEL KARLIN† AND ISTVÁN LADUNGA†‡

†Department of Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Samuel Karlin, August 3, 1994

ABSTRACT **A method for assessing genomic similarity based on relative abundances of short oligonucleotides in large DNA samples is introduced. The method requires neither homologous sequences nor prior sequence alignments. The analysis centers on (*i*) dinucleotide (and tri- and tetra-) relative abundance extremes in genomic sequences, (*ii*) distances between sequences based on all dinucleotide relative abundance values, and (*iii*) a multidimensional partial ordering protocol. The emphasis in this paper is on assessments of general relatedness of genomes as distinguished from phylogenetic reconstructions. Our methods demonstrate that the relative abundance distances almost always differ more for genomic interspecific sequence comparisons than for genomic intraspecific sequence comparisons, indicating congruence over different genome sequence samples. The genomic comparisons are generally concordant with accepted phylogenies among vertebrate and among fungal species sequences. Several unexpected relationships between the major groups of metazoa, fungal, and protist DNA emerge, including the following. (*i*) *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* in dinucleotide relative abundance distances are as similar to each other as human is to bovine. (*ii*) *S. cerevisiae*, although substantially far from, is significantly closer to the vertebrates than are the invertebrates (*Drosophila melanogaster, Bombyx mori*, and *Caenorhabditis elegans*). This phenomenon may suggest variable evolutionary rates during the metazoan radiations and slower changes in the fungal divergences, and/or a polyphyletic origin of metazoa. (*iii*) The genomic sequences of *D. melanogaster* and *Trypanosoma brucei* are strikingly similar. This DNA similarity might be explained by some molecular adaptation of the parasite to its dipteran (tsetse fly) host, a host–parasite gene transfer hypothesis. Robustness of the methods may be due to a genomic signature of dinucleotide relative abundance values reflecting DNA structures related to dinucleotide stacking energies, constraints of DNA curvature, and mechanisms attendant to replication, repair, and recombination.**

Our objective is to describe measures of genomic similarities that do not use (depend on) prior alignment of homologous sequences and to apply them to sufficiently large samples of genomic sequences. Comparisons are based on DNA sequence *relative abundance values* of di-, tri-, and tetranucleotides. These measurements appear to discriminate mostly local genomic structures. Factors that can influence DNA structure include dinucleotide stacking stability, constraints on helicity, and methylation patterns (see *Discussion*). Genomic sequences are analyzed with respect to (*i*) similarities and differences of short oligonucleotide relative abundance *extremes* (1), (*ii*) relative abundance distances within and between genomes, and (*iii*) partial orderings among genomes by comparing the 16-dinucleotide relative abundance values to a set of sequence standards. The approach of this paper departs from almost all other methods of similarity analysis and evolutionary reconstruction by using as its basis sequence information derived from the entire genome rather than individual loci.

DATA AND METHODS

Data. Available nonredundant nucleotide sequences of eukaryotic species from GenBank that in aggregate exceed 100 kb were compiled. These data sets include four fungi, four protists, three invertebrates, and eight representative vertebrates. Most of the species collections exceeded 500 kb. To assess heterogeneity of within- and between-species sequences, the individual sequences were generally combined into samples of 100- to 200-kb lengths (see Table 1).

Symmetrized Frequencies and Relative Abundance Values. Let f_X denote the frequency of the nucleotide X (A, C, G, or T) and f_{XY} denote the frequency of dinucleotide XY. A standard assessment of dinucleotide bias is through the odds ratio: $\rho_{XY} = f_{XY}/f_X f_Y$, which discounts bias in G+C content and general base composition (2). Since DNA structures are influenced by oligonucleotide compositions of both strands (e.g., stacking energies), the formula for ρ_{XY} is modified to accommodate the double-stranded nature of DNA by combining the given sequence and its inverted complement sequence. In this context, the frequency f_A is symmetrized to $f^*_A = f^*_T = (f_A + f_T)/2$ and $f^*_C = f^*_G = (f_C + f_G)/2$. Similarly, $f^*_{GT} = f^*_{AC} = (f_{GT} + f_{AC})/2$ is the symmetrized double-stranded frequency of GT/AC, etc. A symmetrized dinucleotide odds ratio measure is $\rho^*_{GT} = \rho^*_{AC} = f^*_{GT}/f^*_G f^*_T = 2(f_{GT} + f_{AC})/(f_G + f_C)(f_T + f_A)$. The deviation of ρ^*_{GT} from 1 can be construed as a measure of dinucleotide bias of GT/AC.

Relative Abundance Distance Measures. We introduce the *dinucleotide relative abundance distance* between two sequences f and g calculated as $\delta^*(f, g) = (1/16)\Sigma_{ij}|\rho^*_{ij}(f) - \rho^*_{ij}(g)|$, where the sum extends over all dinucleotides (3). Corresponding tri- and tetranucleotide relative abundance distances can be defined controlling for lower-order oligonucleotide biases (3). For a random pair of sequences the ρ^*_{XY} values, for all XY, approach 1 (deviation from 1 is about $1/\sqrt{n}$ for sequences of length n) (4). Therefore for $n \approx 100{,}000$, $|\rho^*_{XY} - 1|$ is of the order 0.003 and the distance between two random sequences is about 0.001. Two hundred simulations produced the range 0.000–0.012.

Genomic Partial Ordering Comparisons. Conceivably, in some cases, the extremes among $\rho^*_{XY}(f)$ and $\rho^*_{XY}(g)$ may dominate the calculations of the distance $\delta^*(f, g)$. To avoid the possibility of a few extreme dinucleotide relative abundances exerting a large influence on the value of $\delta^*(f, g)$, we adapt a method of partial orderings (5). Each sequence is represented by the vector of its 16 dinucleotide relative abundances (ρ^*_{XY}). (It is a matter of indifference whether one regards the relative abundance vector to be of length 16 with all components of weight 1 or of length 10 with 6 nonpalindromic components of weight 2 and 4 palindromic components of weight 1.) The

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡Present address: Department for Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030.

Table 1. Species-specific DNA sequence collections compiled from GenBank

Species (abbreviations)	Total DNA, kb (no. of sequences)	No. of samples formed	Within-species mean distance (range of distances between samples)
Protist			
Dictyostelium discoideum (dis)	488 (425)	3	34 (18–44)
Entamoeba histolytica (his)	111 (86)	1	
Plasmodium falciparum (fal)	833 (1203)	5	21 (12–36)
Trypanosoma brucei (bru)	506 (352)	4	22 (12–28)
Fungus			
Neurospora crassa (cra)	320 (175)	3	11 (9–16)
Aspergillus nidulans (nid)	113 (54)	1	
Saccharomyces cerevisiae (cer)	315 (1)	1*	
	2316 (472)	11†	14 (4–25)
Schizosaccharomyces pombe (pom)	605 (322)	4	11 (7–23)
Invertebrate			
Caenorhabditis elegans (ele)	1036 (1)	10‡	17 (10–26)
	896 (32)	4§	
Drosophila melanogaster (mel)	1420 (2244)	10	15 (6–29)
Bombyx mori (mor)	256 (258)	2	11
Vertebrate			
Trout (tro)	60 (64)	1	
Xenopus laevis (lae)	53 (9)	1§	29 (10–49)
	1382 (1104)	9¶	
Chicken (chi)	136 (9)	1§	22 (12–45)
	2070 (1717)	10¶	
Mouse (mou)	216 (3)	2§	29 (9–57)
	2442 (2283)	10	
Pig	573 (443)	3	18 (11–39)
Rabbit (rab)	168 (9)	1§	28 (9–51)
	1203 (963)	9¶	
Bovine (bov)	110 (9)	1§	25 (8–48)
	1622 (1382)	10¶	
Human (hum)	1205 (21)	10‖	
	1613 (1937)	10¶	35 (4–89)

Samples of 100–200 kb were randomly formed from species-specific sequences. Mitochondrial, rRNA, tRNA, and redundant sequences were excluded. A sample is designated *long* if it is assembled from contigs of length at least 10 kb. Samples composed of sequences <10 kb in length are designated *short*. Samples not designated are composed from sequences varying in length. All numbers of the last column are multiplied by 1000. The average distance is calculated over all sample pairs within each species.

*This sample refers to yeast chromosome III (YCIII).

†Sequences not in YCIII.

‡1-Mb contig (see text).

§*Long* samples.

¶*Short* samples.

‖Samples composed from contigs >30 kb in length.

dinucleotide relative abundance vectors of two genomes are compared with a corresponding 16-component vector of a sequence standard S. If one of the two genomes A and B, say A, is closer to S in at least 13 of the 16 components, a dominance ordering between the two genomes relative to S is determined, expressed as *A dominates B* relative to S. These determinations relative to a standard provide a *partial ordering*. For a given standard, the closest sequences are those that are undominated and dominate several other sequences; the most distant sequences are those that are dominated by several sequences but dominate none. See ref. 5 for applications to herpesvirus molecular evolution.

RESULTS AND ANALYSIS

Intragenomic Homogeneity. The dinucleotide relative abundance distances were calculated between all pairs of sample sequences within species (see Table 1). It is useful to distinguish distance levels such as "random," 0.00–0.015 (see *Data and Methods*), as "very close" (0.015–0.030, e.g., as between bovine and pig), "close" (0.030–0.045, e.g., human to bovine), "moderately related" (0.045–0.065, e.g., frog to mouse), "weakly related" (0.065–0.095, e.g., human to trout), "distantly related" (0.095–0.140, e.g., human to yeast), "distant" (0.140–0.180, e.g., human to *Drosophila*), and "very distant" (≥0.180, e.g., human to *Escherichia coli*, $\delta^* = 0.211$).

Comparisons of 100-kb sections from a 500-kb contig of *E. coli* (data not shown) yield distances in the range 0.008–0.032—i.e., from *random* to *close*. The within-species sample distances for *S. cerevisiae* are strikingly small, from 0.004 to 0.025, and for *S. pombe* samples, from 0.007 to 0.023 (Table 1). Determinations among 100-kb samples obtained by dividing up a 1-Mb contig of *C. elegans* into 10 equal sections yield *random* to *close* mutual distances with average 0.017. Similarly, the insects *D. melanogaster* and *B. mori* within-species sample distances are generally at the level of *very close*. The intragenomic distances for the vertebrate samples produce average values larger and ranges generally more extended than those of the invertebrate and fungal sample sequences. The different samples of human *long* (see legend to Table 1) sequences are more heterogeneous with average within-species distance 0.035 and range 0.004–0.089 (Table 1).

Table 2. Average dinucleotide (upper triangle matrix) and di- plus trinucleotide (lower triangle matrix) relative abundance distances between all DNA samples of the 19 eukaryotic species

	Protist				Fungus				Invertebrate			Vertebrate							
Species	dis	his	fal	bru	cra	nid	cer	pom	ele	mel	mor	tro	lae	chi	mou	rab	pig	bov	hum
dis	—	154	95	122	136	154	110	118	122	135	156	136	126	140	154	142	139	134	130
his	263	—	138	130	70	75	80	107	113	125	150	69	67	88	79	77	84	79	95
fal	170	209	—	142	142	147	111	122	153	115	131	117	103	124	146	125	129	121	116
bru	224	208	205	—	82	102	71	48	78	43	59	111	126	125	173	150	154	148	151
cra	244	173	237	140	—	35	48	62	61	83	94	90	105	113	132	105	115	110	136
nid	251	156	218	143	68	—	62	79	91	95	110	81	119	136	140	124	139	131	153
cer	200	159	174	121	104	99	—	34	69	64	86	79	83	93	124	97	104	96	113
pom	199	184	179	89	113	114	72	—	69	48	62	96	103	106	150	125	130	123	132
ele	203	194	218	129	127	147	112	104	—	76	108	143	135	128	162	129	134	132	154
mel	227	207	188	85	156	158	128	102	131	—	59	121	120	122	172	145	151	143	148
mor	241	227	191	99	156	153	126	95	143	102	—	123	149	156	202	177	182	174	179
tro	237	156	198	166	156	137	128	145	197	172	157	—	55	71	88	71	83	78	92
lae	225	142	174	173	168	166	135	143	185	165	187	94	—	45	60	47	46	40	49
chi	237	165	203	179	182	186	142	151	179	174	194	110	74	—	66	42	40	41	54
mou	254	156	223	227	208	194	173	199	213	224	237	127	96	97	—	49	46	49	58
rab	238	155	203	203	185	181	149	179	185	193	215	110	89	74	78	—	26	30	55
pig	232	162	204	206	190	190	153	180	188	199	217	123	84	69	72	45	—	23	43
bov	228	153	195	198	182	181	143	171	183	191	210	117	74	67	75	54	43	—	41
hum	229	181	196	207	217	214	166	187	207	199	217	133	95	91	91	87	73	74	—

All values are multiplied by 1000. For abbreviations of species names, see Table 1.

Intergenomic Dinucleotide Relative Abundance Distances. These distances are consistent with conventional orderings among vertebrate genomes. Vertebrates are mutually very *close* to *weakly related* and clearly separated from other phyla (Table 2). All vertebrate sample sequences are more similar to any other vertebrate sequence than to almost all nonvertebrate sequences. As expected, the smallest average intergenomic distance was observed for the artiodactyls pig and bovine ($\delta^* = 0.023$) with a lagomorph (rabbit) next closest. The outgroup vertebrate is trout. The distances for trout conform with expectation in that the closest (most similar) vertebrate sequence among those examined corresponds to the amphibian *X. laevis*. The dinucleotide relative abundance distances among fungal species places *A. nidulans* close to *N. crassa* ($\delta^* = 0.035$) and *S. cerevisiae* closest to *S. pombe* (0.034), both smaller than the distance of human to bovine (0.041). The ranges of distances between genomes is given in ref. 3.

We highlight several unexpected findings. (*i*) For each sequence set the fungal sequences (*S. cerevisiae, S. pombe, N. crassa*, and *A. nidulans*), although generally far, tend to be closer to any of the vertebrates than do the invertebrates (*D. melanogaster, B. mori*, and *C. elegans*) (Table 2). (*ii*) To each vertebrate sequence, *S. cerevisiae* is invariably closer than *S. pombe* and to each invertebrate sequence *S. pombe* is closer than *S. cerevisiae*. (*iii*) *D. melanogaster* and *T. brucei* are remarkably closer to each other ($\delta^* = 0.043$) than either is to any of the other eukaryotic sequences; *T. brucei* is also *close* to *S. pombe* (0.048). *B. mori* is generally the most divergent from all vertebrate sequences. The largest distance among all pairings is between *B. mori* and mouse (0.202), almost the same as the distance between human and *E. coli* (0.211). Mostly *D. discoideum* is the farthest from other protist and fungal sequences. Mouse among vertebrates is the farthest from the invertebrates. The above findings prevail for comparisons of the complete DNA species collections as well as for most species samples. For discussions of these DNA relative abundance distances within and between some prokaryote sequences and comparisons between mitochondrial and bacterial genomes, see refs. 3 and 6.

Dinucleotide vs. Dinucleotide plus trinucleotide (di + tri) vs. dinucleotide plus trinucleotide plus tetranucleotide (di + tri + tetra) Distance Correlations. For each reference sequence s and the other genome sequences $g_1, g_2, \ldots, g_{18}$ of Table 1, the array of δ^* distances, $\delta^*(s, g_1), \ldots, \delta^*(s, g_{18})$ was compared with the array of di + tri relative abundance distances $\gamma^*(s, g_1), \ldots, \gamma^*(s, g_{18})$ and with the di + tri + tetra relative abundance distances $\alpha^*(s, g_1), \ldots, \alpha^*(s, g_{18})$. The Pearson and Kendall τ correlations of these arrays are calculated for each species genome and the results are displayed in Table 3. Clearly, the Pearson correlations of di vs. di + tri are almost perfect and di vs. di + tri + tetra are highly correlated. For interpretations of these high correlations, see *Discussion*.

Partial Ordering Comparisons Among Eukaryotes. More sensitive evolutionary relationships ensue from the partial orderings of the 16-component dinucleotide relative abun-

Table 3. Correlation orderings of relative abundance distances

	Reference sequence vs. other genomic sequences		
Species	di vs. di + tri	di vs. di + tri + tetra	di + tri vs. di + tri + tetra
D. discoideum	0.943/0.801	0.874/0.777	0.945/0.924
E. histolytica	0.948/0.753	0.915/0.656	0.984/0.833
P. falciparum	0.868/0.717	0.718/0.572	0.884/0.815
T. brucei	0.954/0.774	0.866/0.682	0.963/0.882
N. crassa	0.952/0.800	0.869/0.712	0.972/0.879
A. nidulans	0.944/0.843	0.854/0.673	0.971/0.830
S. cerevisiae	0.900/0.743	0.772/0.656	0.947/0.854
S. pombe	0.956/0.804	0.887/0.712	0.971/0.908
C. elegans	0.940/0.734	0.830/0.612	0.957/0.866
D. melanogaster	0.944/0.836	0.861/0.682	0.965/0.839
B. mori	0.942/0.783	0.864/0.708	0.972/0.871
Trout	0.876/0.739	0.783/0.678	0.978/0.951
X. laevis	0.942/0.839	0.855/0.697	0.968/0.826
Chicken	0.961/0.774	0.905/0.647	0.979/0.852
Mouse	0.957/0.845	0.893/0.726	0.975/0.875
Rabbit	0.960/0.836	0.921/0.792	0.987/0.944
Pig	0.967/0.832	0.922/0.717	0.984/0.871
Bovine	0.965/0.858	0.915/0.695	0.982/0.849
Human	0.950/0.761	0.897/0.638	0.980/0.826

Pearson/Kendall τ coefficients are calculated for the dinucleotide vs. dinucleotide plus trinucleotide (di + tri) vs. dinucleotide plus trinucleotide plus tetranucleotide (di + tri + tetra) distances from a reference sequence to the other sequences.

dance vectors and reinforce our conclusions from the distance analysis (cf. refs. 3 and 5). For each sequence standard we discuss the strongest dominance orderings (data not shown). The partial orderings are consistent with common evolutionary relationships among the vertebrates. Thus, from any of the bovine, pig, and rabbit standards, the other genomic sequences of this group are undominated and dominate all of the nonvertebrate sequences together with the vertebrate sequences, trout and mouse. From each vertebrate standard, *X. laevis* dominates most of the nonvertebrate sequences. From the amphibian *X. laevis* standard, the avian chicken dominates most other species sequences. From the mouse standard, no vertebrate sequence is dominated. From all vertebrate standards, no nonvertebrate sequence dominates a vertebrate sequence and *B. mori* tends to be the most dominated sequence. From most of the vertebrate standards, human and mouse are often dominated together, suggesting a greater primate–rodent nexus among mammals.

The nearest to *A. nidulans, S. pombe, S. cerevisiae, B. mori*, and *D. melanogaster* dominate the majority of vertebrate sample sequences. With *S. pombe* as standard, *S. cerevisiae* (and *vice versa*) dominates 10 of the 17 other species, indicating markedly that these yeasts are significantly more similar to each other than to any other species in our collection. *D. melanogaster* is dramatically close to the *T. brucei* standard in that *D. melanogaster* dominates 13 of the other 17 species sequences. Among the invertebrate sequences, *C. elegans* appears as an outlier and shows very few dominance relations. With respect to the *B. mori* standard, *D. melanogaster* and *T. brucei* are unexpectedly close. For the protists *D. discoideum* and *P. falciparum* standards, there are few dominance relations, implying substantial divergence from all other eukaryotes considered.

DISCUSSION

In evolutionary relationships of the major groups, such as vertebrates, invertebrates, fungi, plants, and protists, conflicting taxonomic schemes and unresolved issues abound (e.g. refs. 7–9). Conventional approaches base phylogenetic reconstructions on individual genes. However, gene sequence evolutionary comparisons can and often do vary with the gene selected. Our methods address the problem of inferring genomic relationships on the basis of entire genomes, but DNA sequences are not directly compared. Our comparisons within and between species sample sequences are based on *dinucleotide* (di + tri, di + tri + tetra) *relative abundance distances*. The between-species sample distances generally exceed the within-species sample distances, implying robustness of our measure over different parts of the same genome. That within-species distances are smaller than between-species distances seems to indicate that there are factors that impose limits upon compositional relative abundance variation of any particular genome. This notion is supported by the observation of internal consistency for large contigs in *E. coli, S. cerevisiae, C. elegans*, and human (1). Global DNA sequence similarity often is different than protein sequence similarity. Proteins are encoded from polynucleotides putatively less sensitive to local DNA constraints (e.g., stacking, curvature, chromatin) as reflected in dinucleotide relative abundances.

Various questions are raised by our analysis of genomic relative abundance distances. Several recent phylogenetic reconstructions based on both rRNA genes and protein sequences associate fungi to animals more than to plants (e.g., refs. 7 and 8). In this context, compare our finding that places the vertebrate sequences generally more similar to the fungi than to either the protists or invertebrates. As expected the four protist DNA sequence sets are mutually distantly related. However, what can account for the pronounced similarity of *T. brucei* and *D. melanogaster* genomic sequences? Why do the fungal and invertebrate sequences tend to be more random in dinucleotide representations than vertebrate sequences (cf. ref. 1)? We venture some interpretations below.

Vertebrate Evolution. Our dinucleotide relative abundance distances among vertebrates imply orderings consistent with accepted phylogenetic reconstructions. Thus, trout is an outlier among vertebrates (closest to the amphibian *X. laevis*). Mouse is somewhat exceptional among vertebrates, being the farthest from trout or frog. The dominance orderings generated from the artiodactyls (bovine and pig) and lagomorph (rabbit) standards feature trout and mouse among vertebrates as the most dominated by several other vertebrate species.

Divergence of Insects, Nematodes, and Vertebrates. From assessments of dinucleotide relative abundances, most vertebrate collections are significantly distant from those of *C. elegans* and the insects (Table 2). This separation is in agreement with the classic division of most metazoan phyla into two superphyla, Protostoma and Deuterostoma. Somewhat unexpectedly, relative abundance distances of the two superphyla are considerably farther from each other than either of them is from the Ascomycete fungi. The shortest vertebrate–invertebrate dinucleotide relative abundance distance (*X. laevis–D. melanogaster*, 0.120) significantly exceeds the distance between *S. pombe* and *D. melanogaster* collections (0.048). The above observations possibly reflect that genomic structures diverged more rapidly in metazoan branches than in fungal branches or the separate origins of the two superphyla.

Genomic Relations Within Ascomycetes. The mean dinucleotide relative abundance distance between all samples of *S. pombe* and *S. cerevisiae* is as low as 0.034, and their similarity is supported by the partial orderings (see *Results and Analysis*). The mean dinucleotide relative abundance distance among all four major fungal sequence collections (*S. cerevisiae, S. pombe, A. nidulans*, and *N. crassa*) is consonant with the accepted orderings (10). The genomic distances between vertebrates, invertebrates, and the fungi are surprising. They place the fungal species, although far, significantly closer to vertebrates than to invertebrates. This is consistent with the EFI-α protein phylogenetic tree (7). Similarity comparisons of the heat shock protein HSP70, superoxide dismutase, and glutamate dehydrogenase genes also place yeast sequences between vertebrates and invertebrates (details to appear elsewhere).

Protists and Invertebrates. The lowest within-protist distance is relatively high [*D. discoideum–P. falciparum* (0.095), comparable to the human–trout distance of 0.092]. On the other hand, in our collection the highest within-protist distance (*D. discoideum–E. histolytica*, 0.154) is considerably less than the highest within-metazoa distance (*B. mori*–mouse, 0.209).

A surprising result from our analysis is that *T. brucei* DNA is significantly closer to *D. melanogaster* DNA than to any other organism in our study, 0.043 (about the same as between human and bovine), a result also supported by the partial orderings. Taking into account that *T. brucei* spends the larger part of its life cycle in the tsetse fly, which, like *D. melanogaster*, is a dipteran, we may speculate that their genomic closeness may be due to some molecular pathogen-to-host adaptation, coevolution, and/or DNA transfer events between flies and some protists. Along these lines, McClure (11), in her evolutionary studies of reverse transcriptase genes and general retroposons, identified a significant similarity between I and F factors of *D. melanogaster* and the *ingi* element of *T. brucei*.

P. falciparum, like *T. brucei*, is also pathogenic in humans and proliferates in several mosquito vectors including *Aedes aegypti* and *Anopheles* sp. However, unlike *T. brucei*, it is at least distantly related to all metazoan and fungal species examined. *D. discoideum*, as in rRNA and protein studies, seems to be an outgroup substantially divergent from every other species in our collection. *E. histolytica*, a mammalian pathogen, is curiously of weakly related dinucleotide relative abundance distance to vertebrates (e.g., *E. histolytica–X. laevis*, 0.067), comparable to the chicken–mouse distance (0.066).

What Do Relative Abundance Distances Measure? The dinucleotide relative abundance distance measure between DNA sequences appear to provide meaningful comparisons (e.g., ref. 5). We suggest that the short oligonucleotide (di-, tri-, and tetranucleotide) relative abundance values relate to DNA structures. Several factors that influence DNA structures have been identified—e.g., dinucleotide stacking energies, curvature, superhelicity, methylation and other short oligonucleotide modifications, and DNA repair mechanisms (12–17). For example, TpA is intrinsically less stable energetically than all other dinucleotides (12, 13). Flexibility of the TpA step is commonly associated with substantial DNA distortions. Untwisting and bending at TpA sites occurs in transcription initiation via protein binding to the TATA box, *Eco*RV binding to its recognition sequence, and $\gamma\delta$ resolvase binding to the site at which crossing-over occurs (15, 17). These models suggest that TpA sites can be important as nucleation sites for untwisting the DNA double helix. It appears that protein–DNA complexes can exploit the reduced thermodynamic stability of the TpA base step. The TpA and ApT steps are conformationally incompatible causing a strain in the helix when juxtaposed, which can be relieved by unwinding the helix (16).

Hunter (16) set forth a theoretical framework for understanding and predicting the sequence-dependent structure and properties of double-stranded DNA. The analysis is based primarily on the energetics of base stacking interactions. These take account of (*i*) cross-strand steric clashes, for example, at pyrimidine–purine steps, especially collision between the thymine methyl group and the 5′ neighboring sugar (which causes a negative propeller twist), and of (*ii*) electrostatic interactions between partial atomic charges and the π electrons of the aromatic rings.

Dinucleotide relative abundances capture most of the departure from randomness in DNA sequences. Comparisons were made in terms of di + tri (and di + tri + tetra) relative abundance distances. The di and the corresponding di + tri (and the di + tri + tetra) relative abundance distances between sequences highly correlate (Table 3), suggesting that DNA conformational stacking arrangements are principally determined by base-step configurations. Observation of the distribution of dinucleotides separated by 0, 1, or 2 other nucleotides has shown that, although values for 0 space are highly biased, those for space 1 or 2 are more nearly random (18). Furthermore, a theoretical investigation of the energy minima for the geometry of two neighboring base pairs in terms of slide, roll, and helical twist parameters finds that the 16 dinucleotides largely provide the DNA structures observed with x-ray diffraction of synthesized oligonucleotides (16).

DNA has at least two functions: (*i*) to provide special sequences for encoding gene products or for regulating transcription and (*ii*) to provide for genome replication and segregation. While the former requires some sequence specificity, the latter may be mostly DNA structure specific. Cell divisions involve DNA stacking on itself that needs to be appropriately decondensed to undergo segregation. In higher eukaryotes, controls on replication and segregation are not understood and origins of replication may not be strictly sequence specific but perhaps more structure specific (19). The genome putatively requires compositional flexibility and balance and conveys controls and information in terms of both DNA structure and sequence. In this vein, the relative abundance distances appear to assess and discriminate mostly local structure specificity, whose signature (the symmetrized 10-component dinucleotide relative abundance values) is well conserved within genomes (18).

We gratefully acknowledge discussions on the manuscript with Drs. B. E. Blaisdell, V. Brendel, P. Bucher, A. M. Campbell, R. F. Smith, G. Vida, G. Weinstock, and K. C. Worley. This research was supported in part by National Institutes of Health Grants 2R01HG00335-06 and 5R01GM10452-30 and National Science Foundation Grant DMS91-06974.

1. Karlin, S., Ladunga, I. & Blaisdell, B. E. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 12837–12841.
2. Burge, C., Campbell, A. M. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 1358–1362.
3. Karlin, S. & Cardon, L. R. (1994) *Annu. Rev. Microbiol.* **48**, 619–654.
4. Hollander, M. & Wolfe, D. A. (1973) *Nonparametric Statistical Methods* (Wiley, New York).
5. Karlin, S., Mocarski, E. S. & Schachtel, G. A. (1994) *J. Virol.* **68**, 1886–1902.
6. Karlin, S. & Campbell, A. M. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 12842–12846.
7. Baldauf, S. L. & Palmer, J. D. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 11558–11562.
8. Wainright, P. O., Hinkle, G., Sogin, M. L. & Sticket, S. K. (1993) *Science* **260**, 340–342.
9. Novacek, M. J. (1992) *Nature (London)* **356**, 121–125.
10. Bruns, T. D., White, T. J. & Taylor, J. W. (1991) *Annu. Rev. Ecol. Syst.* **22**, 525–564.
11. McClure, M. A. (1991) *Mol. Biol. Evol.* **8**, 835–856.
12. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 3746–3750.
13. Delcourt, S. G. & Blake, R. D. (1990) *J. Biol. Chem.* **266**, 15160–15169.
14. Calladine, C. R. & Drew, H. R. (1992) *Understanding DNA* (Academic, San Diego).
15. Travers, A. A. & Schwabe, J. W. R. (1993) *Curr. Biol.* **3**, 898–900.
16. Hunter, C. A. (1993) *J. Mol. Biol.* **230**, 1025–1054.
17. Travers, A. A. (1993) *DNA-Protein Interactions* (Chapman & Hall, London).
18. Karlin, S. & Burge, C. (1995) *Trends Microb. Sci.*, in press.
19. Krysan, P. J., Smith, J. G. & Calos, M. P. (1993) *Mol. Cell. Biol.* **13**, 2688–2696.