6228–6236 *Nucleic Acids Research, 2008, Vol. 36, No. 19* *Published online 1 October 2008*
doi:10.1093/nar/gkn626

A novel DNA sequence periodicity decodes nucleosome positioning

Kaifu Chen[1,2], Qingshu Meng[1,2], Lina Ma[1,2], Qingyou Liu[3], Petrus Tang[4], Chungshung Chiu[5], Songnian Hu[1] and Jun Yu[1,*]

[1]Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, [2]Graduate University of Chinese Academy of Sciences, Beijing, [3]Animal Reproduction Institute, Guangxi Key laboratory of subtropical Bioresource Conservation and Vitalization, Guangxi University, Nanning, [4]Bioinformatics Center/Molecular Medicine Research Center, Chang Gung University, Taoyuan 333, Taiwan and [5]Molecular Infectious Diseases Research Center, Chang Gung Memorial Hospital, Taoyuan 333, Taiwan, China

Received June 16, 2008; Revised August 16, 2008; Accepted September 12, 2008

ABSTRACT

There have been two types of well-characterized DNA sequence periodicities; both are found to be associated with important molecular mechanisms. One is a 3-nt periodicity corresponding to codon triplets, the other is a 10.5-nt periodicity related to the structure of DNA helixes. In the process of analyzing the genome and transcriptome of *Trichomonas vaginalis*, we observed a 120.9-nt periodicity along DNA sequences. Different from the 3- and 10.5-nt periodicities, this novel periodicity originates near the 5′-end of transcripts, extends along the direction of transcription, and weakens gradually along transcripts. As a result, codon usage as well as amino acid composition is constrained by this periodicity. Similar periodicities were also identified in other organisms, but with variable length associated with the length of nucleosome units. We validated this association experimentally in *T. vaginalis*, and demonstrated that the periodicity manifests nucleotide variations between linker-DNA and wrapping-DNA along nucleosome array. We conclude that this novel DNA sequence periodicity is a signature of nucleosome organization suggesting that nucleosomes are well-positioned with regularity, especially near the 5′-end of transcripts.

INTRODUCTION

DNA sequence variations are created through synthetic errors, concerning two essential molecular mechanisms: genome replication and DNA damage–repair; these errors are fixed gradually among genomes through drift and selection over time, leaving various sequence signatures (1–3). DNA sequence periodicity represents one of such signatures, and there have been two well-characterized periodicities among genomes with either 10.5 nt or 3 nt in periodicity length. The 10.5-nt periodicity tends to be close to 11 nt in bacteria, but it is close to 10 nt in archaea and eukaryotes (4). The plausible explanation for the 10.5-nt periodicity has been attributed to chromatin organization (5,6), where DNA double helix exhibits a periodicity of approximately 10.5 nt (7,8), and among eukaryotic genomes it is further folded into nucleosome arrays (9–11). Each nucleosome is composed of a histone octamer wrapped around by double-stranded DNA (12,13), and this packaging makes bases face inward toward the histone octamer less accessible than those facing outward, resulting in a 10.5-nt periodicity in nucleotide composition (5,6,9). The longer length of this periodicity in bacteria was postulated to reflect a less condensed organization of prokaryotic chromatins, and was suggested to be associated with negative DNA supercoiling (14). The 3-nt periodicity is characteristic of protein-coding sequences, and perhaps arises from the selective pressure for proteins and their codons (15–17).

We started our investigation on DNA sequence periodicity of *Trichomonas vaginalis*, a unicellular parasite found in human urogenital tracts (18,19), as its genomic DNA degrades unusually fast when a common procedure for isolating large DNA fragments were carried out. This organism is an ancient protozoa that lacks mitochondrion but contains hydrogenosome instead, albeit rooted deeply in the eukaryotic branch (19). Recent studies revealed that *T. vaginalis* genes possess unusually short 5′- and 3′-UTR, 5–20 nt in length (20), and that the organism's genome harbors a large number of heavily duplicated genes (21). In this work, we first describe the discovery of a novel sequence periodicity and its characteristic in the *T. vaginalis* sequence, then generalize this observation among

*To whom correspondence should be addressed. Tel: +86 10 8299 5357; Fax: +86 10 8299 5373; Email: junyu@big.ac.cn other unicellular organisms, and finally discuss the biological implications of this novel periodicity.

MATERIALS AND METHODS

The datasets

Our *T. vaginalis* expressed sequence tag (EST) data includes 71 594 EST sequences generated in our laboratories (Petrus Tang *et al.*, unpublished data) and 20 002 additional ESTs from GenBank. The *T. vaginalis* genome sequence (~177 Mb) and 99 319 predicted genes were downloaded from the TIGR database (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_vaginalis/) on 24 January 2007. The sequence data for assessing nucleosome binding in *Saccharomyces cerevisiae* were kindly provided by Vishwanath R. Iyer (http://www.iyerlat.org/nucleosome). All other datasets used for this analysis were from GenBank databases.

To separate *T. vaginalis* genes from the intergenic sequences, we aligned ESTs to the assembled genomic sequences, predicted genes and classified the predicted genes into three categories according to reliability. First, if an EST matches to a genomic region that corresponds to an annotated gene, and if its alignment score is 300 above the second highest score for the same EST, we considered this gene as verified by this EST. This definition yielded a set of 8093 most reliable genes. Second, if a predicted gene matches EST sequences with a score lower than 100, or no EST matches, but yielding clear annotations, we classified these genes as moderately reliable; this dataset contains 6712 genes. Third, when genes are annotated as 'hypothetical proteins' or did not yield any meaningful information from their annotations, we classified them as the least reliable gene set that grouped 47 812 genes. We also aligned all the predicted genes to the genome, and masked all matched genomic sequences with a blast score of 30 or above; this process yielded 36.67 Mb unmasked fragments or intergenic sequences, which are longer than 200 bp.

Power spectra

Power spectrum analysis is a popular method for detecting periodicity in numerical sequences. To accelerate calculation, we used Fast Fourier Transform algorithm to compute power spectrum. For a sequence x_k of length $2N$ (N is a positive integer), its power spectrum is expressed as:

$$S(f_j) = \left| \sum_{k=1}^{2N} x_k \exp(-2\pi i_k f_j) \right|^2$$

where $i^2 = -1$, and $f_j = j/2N$ ($j = 0, 1, \ldots, N$).

To apply power spectrum analysis to DNA sequences, we first translated each DNA sequence into a binary sequence x_k: when nucleotide G or C present at position k, $x_k = 1$, otherwise, $x_k = 0$. Resultant binary sequences are then joined into one sequence, and again split into $2N$-nt fragments ($N = 8192$ in Figure 1a; $N = 512$ in Figures 1b and 3). Power spectrum was calculated for each of these fragments first and plotted the results for all fragments after being averaged.

To distinguish periodicities that are in phase with CDS-start, we have applied power spectrum analysis in a novel way. After aligning transcripts with the origin at the junction of the 5′-UTR and the coding sequence (CDS), we calculated GC content (X) for each nucleotide position (K) in bulk for transcripts, generated a binary sequence X_k ($k = 0, 1, \ldots, 1024$), and subsequently applied the power spectrum analysis to the binary sequence.

Micrococcal nuclease digestion

We harvested *T. vaginalis* (To16) cells (20 mg) from a 36-h culture by centrifugation and, washed twice with 200 µl PBS (pH 6.0), and resuspended in 400 µl digestion buffer (20 mM Tris–HCL, 5 mM NaCl, 2.5 mM $CaCl_2$, pH 8.0). The cells were homogenized with a PRO200 homogenizer for 30 s. The lysates were incubated with 0.5 µl (~10U) micrococcal nuclease for 1, 3, 5, 7, 9, 11,

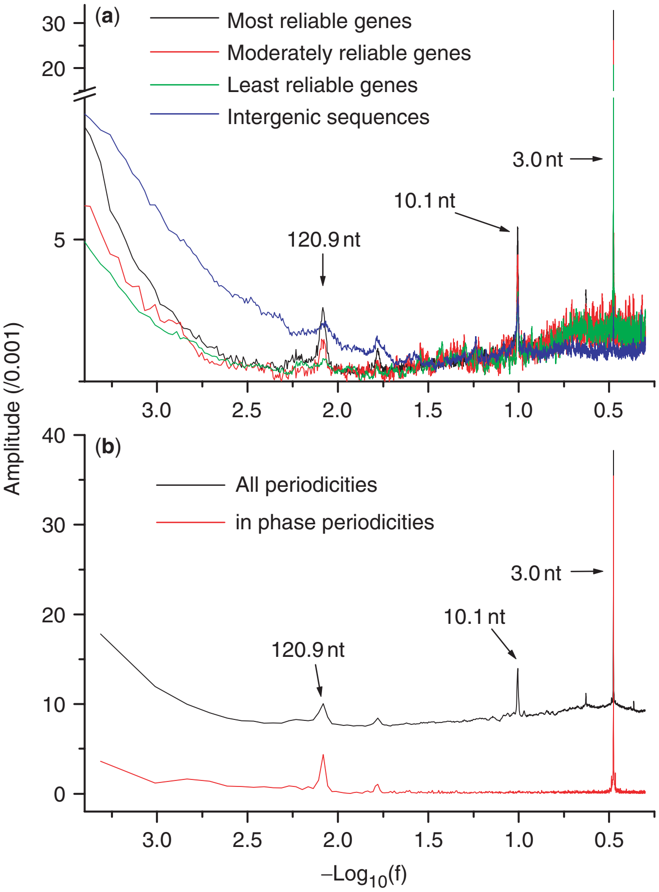

Figure 1. Power spectrum density (PSD) analysis of *T. vaginalis* genomic sequences. Periodicities correspond to spikes above the baseline in the plots. (**a**) When PSD was applied in a traditional way to identify all sequence periodicities in four different datasets (most reliable genes, moderately reliable genes, least reliable genes and intergenic sequences), there were three major spikes found at length ranges of 3.0, 120.9 and 10.1nt. Periodicities appear more pronounced in more reliable genes. (**b**) When PSD was applied in a novel way to distinguish periodicities that are in phase with CDS-start, only 3.0 and 120.9-nt periodicities are identified as in phase with CDS-start.

13, 15, 17, 19 and 21 min at 37°C, and the reactions were stopped with 25 mM EDTA. The reaction mixtures were subsequently incubated with 12 µg pancreatic RNase for 1 h at 37°C, followed by incubation with 0.5% SDS for another hour at 37°C and 80 µg proteinase K at 50°C for 3.5 h. Cold ethanol (2.5 volumes of the mixture) was added to precipitate DNA. The isolated DNA was resuspended in 20 µl TE, and 10 µl of this final DNA solution was subjected to electrophoresis with 3% agarose gels.

RESULTS

Power spectrum analysis revealed two types of nucleotide periodicities

Our power spectrum analyses for DNA sequence periodicity began with the *T. vaginalis* genome sequence based on four datasets: the most reliable, moderately reliable, and least reliable gene sets as well as a collection of intergenic sequences. Three major peaks at 120.9, 10.1 and 3.0 nt in unit lengths were readily observed (Figure 1). The 3- and 10.1-nt periodicities have been reported previously in genomes of both prokaryotes and eukaryotes (22), whereas the 120.9-nt periodicity have not yet been described in literature. All these three periodicities appear more pronounced for gene sequences (Figure 1a), suggesting potential associations with genic or protein-coding sequences. To further investigate these associations, we improved the algorithm for power spectrum analysis to discriminate periodicities that are in phase with CDS-start or not (Figure 1b). In-phase periodicities should share same phase among genes at the junction of 5′-UTR and CDS. The two major peaks, at the coordinates of 120.9 and 3 nt, remained in the spectrum for in-phase periodicities but the 10.1-nt periodicity disappeared (Figure 1b), since the 3-nt periodicity is a signature of the codon triplet (15,16) and the 10.1-nt periodicity is associated with DNA double-helix structure and the eukaryotic chromatin organization (5,6,14).

The 120.9-nt periodicity in the *T. vaginalis* genic sequences

To further investigate the 120.9-nt periodicity, we calculated GC content for each nucleotide position along transcripts among the most reliable set of genes, partitioning the three nucleotide positions of codons into GC1, GC2 and GC3 (Figure 2a). The 120.9-nt periodicity showed several interesting characteristics. First, it is most pronounced at the third codon position (Figure 2a and b), where at least half of the nucleotide variations do not lead to amino acid composition changes (23). The periodicity also appeared stronger at GC1 than GC2, largely due to the rigidity at the second codon position where nucleotide variations often lead to amino acid changes that convert physiochemical properties (23). Second, the periodicity appears starting at the junction between 5′-UTR and CDS. To further validate this observation, we plotted the GC content with origins at the junction between CDS and 3′-UTR, then the 120.9-nt periodicity was abolished (Figure 2c). We also plotted the GC content starting at the second ATG codon downstream of the real (first) start codon, and observed a weakening periodicity (Figure 2d). Third, the 120.9-nt periodicity seems more striking at the 5′-end of CDS, extends only along the

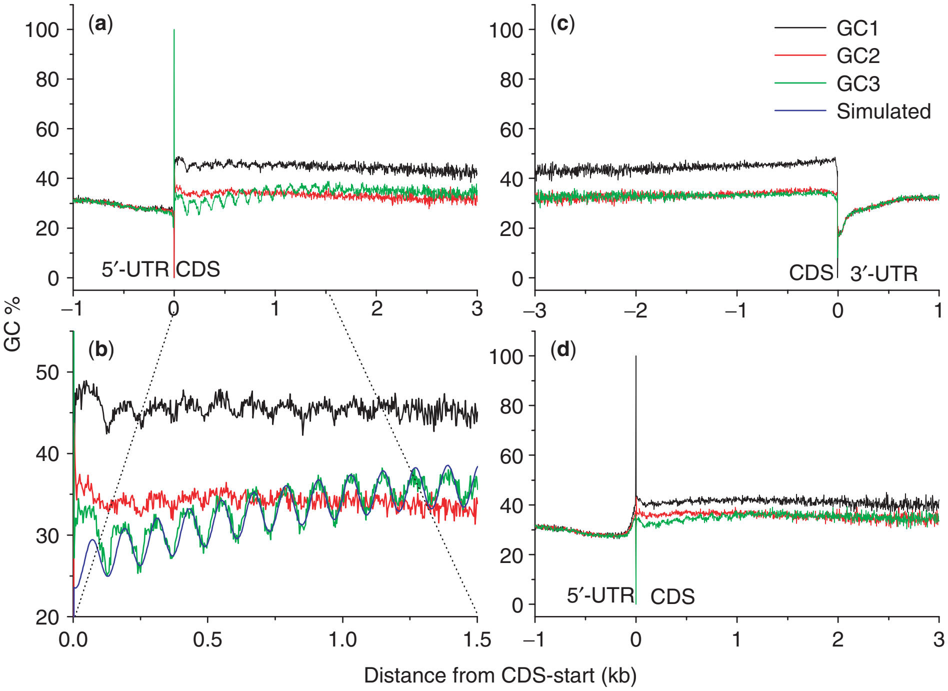

Figure 2. GC-content plotted as a function of nucleotide position among genes. (**a**) The plot was generated with origin at the junction between 5′-untranslated and coding regions. (**b**) An enlarged view of what in (a). (**c**) The plot was generated with origin at the junction between coding and 3′-UTRs. (**d**) The plot was generated using the second ATG codon instead of the real start codon as the origin. The value at each position was an average over the most reliable genes. Three nucleotide positions for codon triplets were plotted separately as GC1, GC2 and GC3.

direction of transcription, and deteriorates completely beyond ~2000 nt. To exclude the possibility of any length effect, we divided our collection of genes into six different length classes, and did not observed obvious deviations from the norm (Supplementary Figure S1). Fourth, despite the fact that this periodicity seemed aligning with the translation start site, we believe that it initiates from the transcription start site for two reasons. One points to the extremely short 5′-UTR in the *T. vaginalis* genes so that we are unable to distinguish the difference between these two starts. The other is the uneven selection pressure between protein-coding and nonprotein-coding sequences, so that the sequence signature may not be detectable in the UTR regions when compared to protein-coding regions.

In addition to the periodicities, we also observed a positive GC-content gradient along genes. Similar to the case of the 120.9-nt periodicity, the gradient effect appears more pronounced in GC3 and at the 5′-end of CDS. The GC3 curve was fitted as:

$$X_{(k)} = [-32(K/30000)^3 - 32(K/30000)^2 + 3.7K/30000 + 0.26] \times 100 + [2.6\sin(2\pi(K-40)/120.9)]$$

where $X_{(k)}$ represents GC content at a position that is K nucleotides away from the junction between 5′-UTR and CDS. The right side of this equation can be split into two parts: the first

$$[-32(K/30000)^3 - 32(K/30000)^2 + 3.7K/30000 + 0.26] \times 100$$

describes a baseline for GC3, and the second

$$[2.6\sin(2\pi(K-40)/120.9)]$$

yields a sine undulation with a period length of 120.9 nt.

The universal 10.5-nt periodicity in genomes of all three kingdoms was explained primarily as a periodic distribution of dinucleotides, and different dinucleotide contributes differently to the periodicity, such as the dominant effect of ApA/TpT (22). Although the individual nucleotides contribute to the 120.9-nt periodicity with little difference (Supplementary Figure S2a), the 120.9-nt periodicity seems to prefer different combinations of the 16 dinucleotide, ranging from the strongest, such as ApA and TpT, to the weakest, such as TpA, GpA and CpG (Supplementary Figure S2b). In addition, this periodicity has obvious directionality along the direction of transcription, manifesting clearly among the complementary dinucleotide pairs, such as ApC/GpT and CpA/TpG.

Universality of the 120.9-nt periodicity

The ability to detect a transcript-centric periodicity is governed by strength of the periodicity relevant to reliability and adequacy of the datasets. Since the 120.9-nt periodicity in *T. vaginalis* genome is rather transcript-centric, the amount of EST-validated gene sequences played a critical role in its successful detection (Figure 3). We have tried to collect data from other organisms and attempted to generalize this discovery with a caveat of lacking adequate amount of validated gene sequences and pronounced effects. We did observe similar periodicities in some of the organisms simply by applying our power spectrum analysis, especially in some archaea and unicellular eukaryotic species (Figure 3). To further validate the results, we also tried to illustrate these periodicities in mononucleotide and dinucleotide compositions, as we have done for the 120.9-nt periodicity in *T. vaginalis* genome, and observed at least two similar periodicities with lengths of ~132 and 156 nt in *Giardia lamblia* and *Yarrowia lipolytica*, respectively (Supplementary Figure S3). Similar to the *T. vaginalis* 120.9-nt periodicity, these periodicities also start at the 5′-end of CDS, appeared more obvious in GC3 along the direction of transcription, extending over 1500 nt in length. The dinucleotide preference of the two periodicities is distinguished from that of the *T. vaginalis* 120.9-nt periodicity, in that the 132-nt periodicity of *G. lamblia* prefers ApT, TpA and CpC, whereas the 156-nt periodicity of *Y. lipolytica* prefers GpC, CpG and ApG.

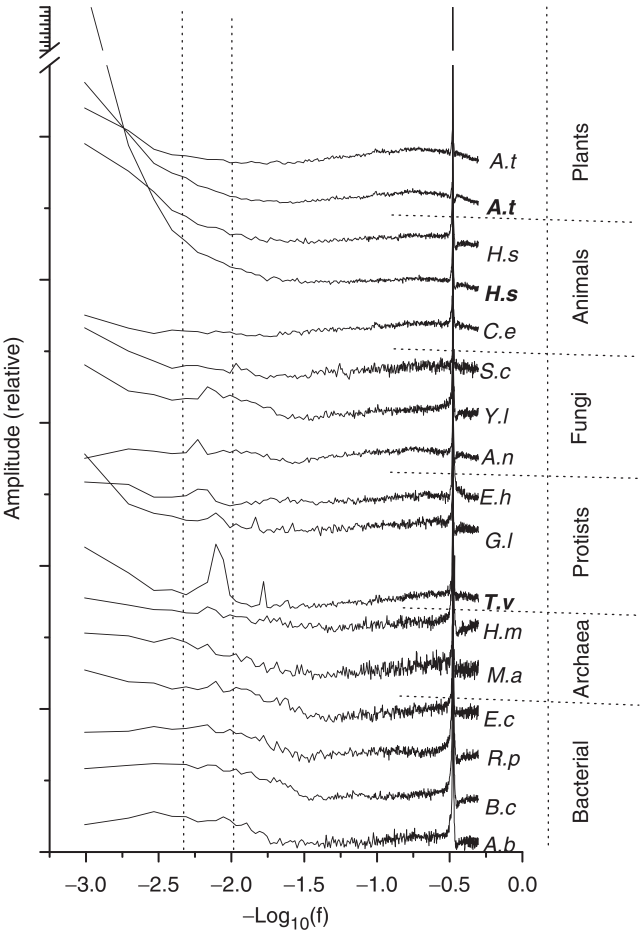

Figure 3. Power spectra for periodicities that are in phase with CDS-start in genes of different species. Most of the spectra were based on coding sequences from selected multicellular and unicellular organisms, except those labeled in bold—which were EST-verified coding sequences of *T. vaginalis* and transcript sequences that include multiple introns and exons of *Arabidopsis thaliana* and *Homo sapiens*.

On the biology implication of the 120.9-nt periodicity

In the process of identifying similar periodicities among different species, we observed that the lengths of these periodicities range from 150 to 200 nt, which is capable of folding into a nucleosome unit in eukaryotes (24). We realized that these periodicities may relate to the organization of nucleosomes among different species as histone polymers in archaea and eukaryotes form stable complex with 50–200 bp DNA sequences (25–28). We went on to measure the unit length of *T. vaginalis* nucleosomes based on micrococcal nuclease digestion (Figure 4). The DNA ladders with fragment lengths of 480, 360 and 240 bp suggested the existence of a repetitive unit protected by nucleosome structures with a length of ~120 bp. This length is equivalent to the length of the novel periodicity although a precise size measurement with single-basepair resolution is impossible, largely limited by the rather poor resolution of agarose gel electrophoresis and the dynamic nature of enzyme digestions. The 100 bp and 140 bp bands are believed to be nucleosome-associated DNA fragments without any and with the two linker sequences, respectively.

Another empirical evidence for the nucleosome relevance of this periodicity came from a study on nucleosome fingerprinting of *S. cerevisiae*. Although its periodic behavior in genomic GC content appeared much weaker (albeit better at GC3) as compared to that of *T. vaginalis* genome (Figure 5a), a high-resolution map on nucleosome positioning provided us an adequate amount of nucleosome-protected DNA sequences to interrogate how nucleotide composition is constrained by nucleosome organization (29). When aligned the nucleosome-protected sequences to calculate GC content along nucleosome arrays (Figure 5b), we observed that the nucleosome-protected DNA is more GC-rich than the linker DNA, and this local GC content varies with a periodicity corresponding to its nucleosome unit length.

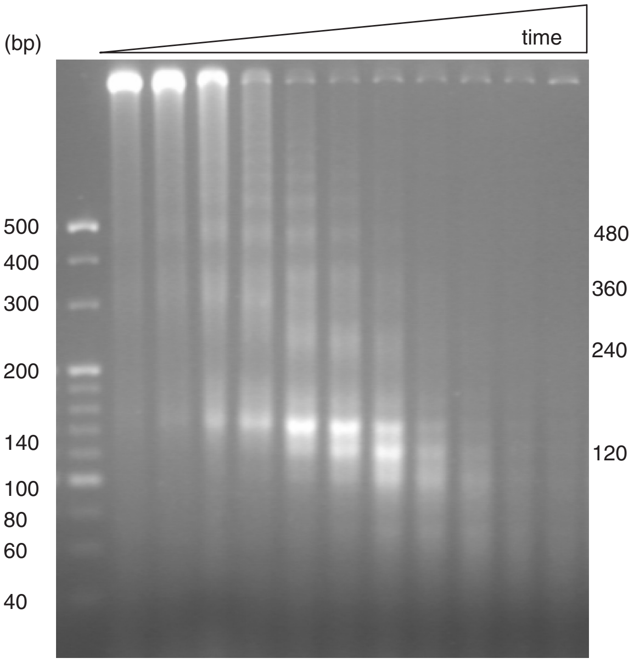

Figure 4. Electrophoretic analysis of *T. vaginalis* chromosomal DNA fragments protected by nucleosomes based on micrococcal nuclease digestion. Time-controlled enzyme digestions were performed for 1, 3, 5, 7, 9, 11, 13, 15, 17, 19 and 21 min at 37 °C. Bands identified at 480, 360 and 240 bp suggested the existence of a repetitive unit with a length of ~120 bp. The 140- and 100-bp bands are believed to be nucleosome-associated DNA fragments with and without two linker sequences, respectively.

Codon usage and amino acid composition

Codon usage can be biased among genes and genomes toward GC or purine contents (23,30–34). A codon-based analysis named effective number of codons (*Ne*) was chosen to quantify codon usage bias along CDS (35–37). We calculated *Ne* at each codon position for all genes aligned from the translation initiation site (TIS; Figure 6a). *Ne* usually ranges from 20 (when only one codon is used for each amino acid) to 61 (when there is

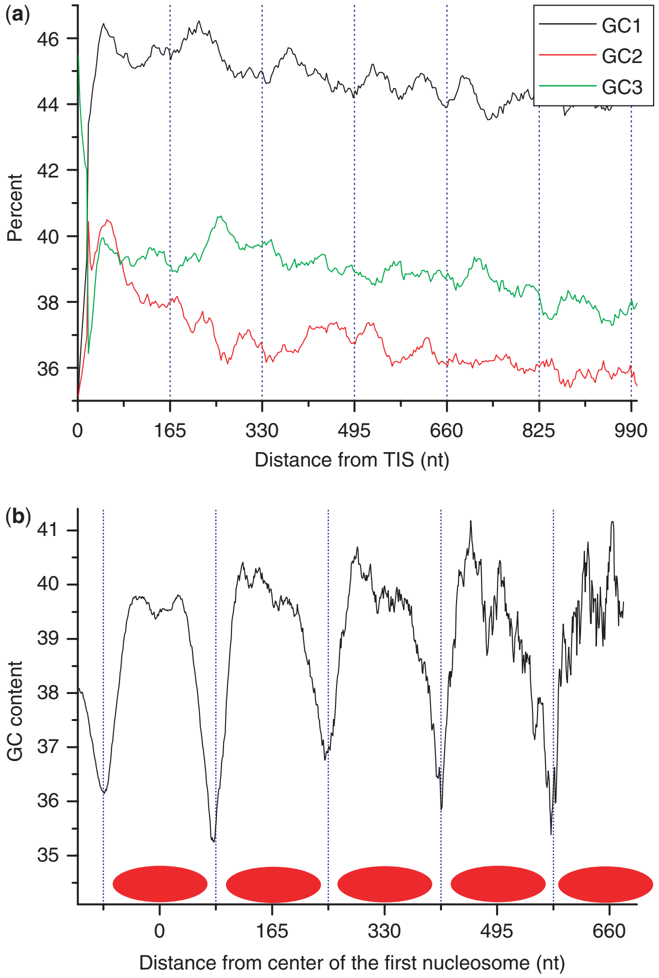

Figure 5. The 165-nt sequence periodicity in *S. cerevisiae*. (**a**) GC content at three codon positions were plotted separately as GC1, GC2 and GC3 along coding DNA sequences. (**b**) Nucleosome-binding DNA sequences were aligned from one end to calculate GC content along nucleosome array; each red oval represent one nucleosome position. A 10-nt sliding window was used to smooth the curves.

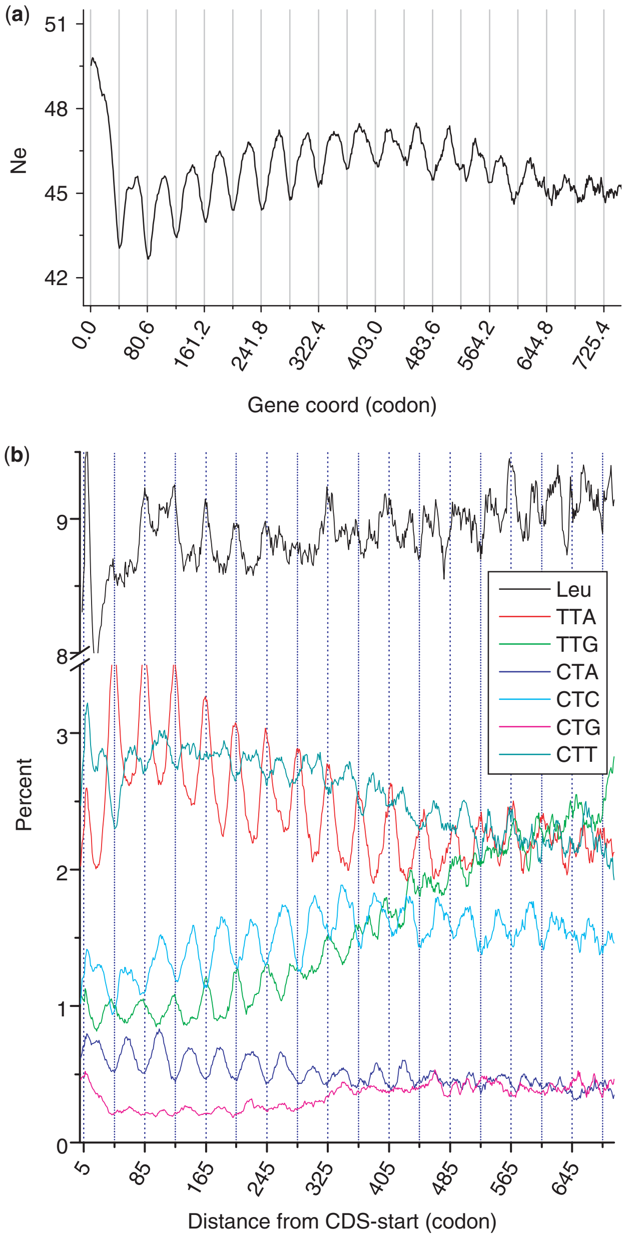

Figure 6. Codon usage analyses over aligned coding sequences. (**a**) Effective number of codons was plotted as a function of codon positions along the direction of translation. (**b**) The 120.9-nt periodicity in the six leucine codons. Value at each position was averaged among the most reliable gene set of *T. vaginalis* and a sliding window of 10 codons was used to smooth the curves. Note that periodicities are more obvious when certain codons were plotted individually albeit poor appearance at amino acid level.

no bias at all). We observed that *Ne* fluctuates with the 120.9-nt (40.3 codons) periodicity, suggesting that codon usage is significantly constrained by this periodicity (Figure 6a). In the case of leucine, four of its six codons show an obvious periodicity except the CTY codons—CTT and CTG—which just appeared noisier (Figure 6b). Another feature is that the periodicities of different codons appeared to fall into two phases, either in-phase or in-reverse-phase with each other. We thus went on to scrutinize all codons of 20 amino acids and found several rules. First, G- and C-ending codons are generally in the same phase (valley), whereas A- and T-ending codons are in another (peak). Second, although there are 12 exceptions to the first rule, all amino acids with 2-fold degenerate codons follow the first rule firmly. The two degenerate codons for each amino acid are either purine- or pyrimidine-ending, remaining in the opposite phases (peak versus valley). Third, all amino acids with three or more codons have at least one exception to the first rule. The amino acids with three and four codons have only one exception for each and those with six codons have one to three exceptions depending on their usages. For instance, the most abundant amino acid is leucine in all genomes so it has three exceptions, as compared to serine and arginine, which have two and one exceptions, respectively. Fourth, all exceptional codons are either A- or T-ending except one codon for the most abundant amino acid—leucine, which is a G-ending codon. It becomes obvious that these rules are all pointing to a biased codon usage pressured by mutations in favorite of GC-decrease or AT increase from 5′ to 3′ along the transcripts. Further more, the periodicity and the phase for amino acids are governed by a collective effect of all their codons.

DISCUSSION

Intensive computational analyses allowed us to identify a novel 120.9-nt periodicity in *T. vaginalis* genome and generalize its existence in several other organisms, where the periodicity length varies in accordance with the unit length of their nucleosome DNA. This length association, along with the validation from other newly available data type—*S. cerevisiae* nucleosome-binding DNA sequences, reveals that this novel periodicity exhibits different nucleotide compositional dynamics between linker-DNA and wrapping-DNA of nucleosomes. The preference of this periodicity toward CDS, and its start at the junction between 5′-UTR and CDS, along with its extension along the direction of transcription, suggest that nucleosomes are well-positioned near 5′-end along CDS. This is in agreement with several recent reports demonstrating that the coding region downstream of the start codon (or transcription start site) is well-organized into nucleosome arrays, and there is usually a nucleosome-free region upstream of the start codon (29, 38–41).

The nucleosome-associated nature of the periodicity suggested that its characteristics should be universal to all eukaryotic genomes. However, there are several reasons why we only succeeded in discovering the 120.9-nt periodicity in *T. vaginalis* and a few other genomes, but are unable to identify the comparable periodicity in the majority of other species. First, the strength of the transcript-centric periodicity is associated with reliable identification of genes in the genome and correct alignment of their transcripts. The great amount (more than 60 000 genes and 10 000 of them were EST-verified) of annotated genes in *T. vaginalis* genome provided adequate materials for the detection as we have not seen enough data among other genomes, especially other unicellular organisms (Tables 1 and 2). Second, since the periodicity

Table 1. Comparison between genomes of species investigated in this work

Species	Kingdom	Genome size	GC (%)	CDS number	CDS with intron (%)
Arabidopsis thaliana	Plant	119.19	36	30 480	80.02
Homo sapiens	Protist	3230.8	41	26 336	85.20
Caenorhabditis elegans	Protist	100.2	35	22 844	97.45
Saccharomyces cerevisiae	Fungus	12.16	38	5877	5.36
Yarrowia lipolytica	Fungus	20.5	49	6520	10.28
Aspergillus nidulans	Fungus	31	50	9396	N/A
Entamoeba histolytica	Protist	22.86	25	9772	24.99
Giardia lamblia	Protist	9.72	48	6569	0.02
Trichomonas vaginalis	Protist	177.89	33	99 319	0.07
Haloarcula marismortui	Archaeon	4.27	61	4240	0.00
Methanosarcina acetivorans	Archaeon	5.75	43	4540	0.00
Escherichia coli	Bacterium	4.64	51	4243	0.02
Rhodopseudomonas palustris HaA2	Bacterium	5.33	66	4683	0.00
Burkholderia sp. 383	Bacterium	8.68	66	7717	0.00
Acidobacteria bacterium Ellin345	Bacterium	5.65	58	4777	0.00

N/A, data not available.

Table 2. The 120.9-nt periodicity in codons and amino acids of *T. vaginalis*

Codon	%	Strength	Phase	AA	%	Strength	Phase	Codon	%	Strength	Phase	AA	%	Strength	Phase
AAA	4.48	+++	P	K	7.62	++	N	ACA	3.15	+++	V	T	5.49	++	V
AAG	3.14	+++	V					ACC	0.49	+	V				
AAC	2.48	+++	V	N	6.2	++	P	ACG	0.4	+	V				
AAT	3.71	+++	P					ACT	1.45	++	P				
ATG	2	++	V	M	2	++	V	TGG	0.77	+	V	W	0.77	+	V
ATA	1.53	+++	V	I	8.02	++	P	TGC	0.96	+	V	C	1.63	+	N
ATC	2.5	++	V					TGT	0.67	+	P				
ATT	3.99	+++	P					AGC	0.66	++	V	S	7.82	++	N
TAC	0.17	++	V	Y	3.59	+	N	AGT	0.94	++	V				
TAT	1.93	++	P					TCA	2.67	++	V				
TTC	2.87	++	V	F	5.15	++	P	TCC	1	++	V				
TTT	2.29	+++	P					TCG	0.55	+	V				
GAA	5.53	+++	P	E	7.21	++	N	TCT	2	++	P				
GAG	1.65	+++	V					GGA	1.83	+	P	G	4.21	++	V
GAC	1.65	+++	V	D	5.82	++	V	GGC	1.03	+	V				
GAT	4.17	++	P					GGG	0.11	+	V				
GTA	0.97	++	V	V	5.33	++	V	GGT	1.24	++	V				
GTC	1.24	++	V					GCA	2.23	++	V	A	5.73	++	V
GTG	0.29	+	V					GCC	0.96	+	V				
GTT	2.82	++	P					GCG	0.37	+	V				
CAA	2.82	+++	P	Q	4.29	++	N	GCT	2.17	++	P				
CAG	1.47	++	V					AGA	1.89	+	P	R	3.69	++	V
CAC	0.68	+	V	H	1.85	++	N	AGG	0.38	+	V				
CAT	1.16	+	N					CGA	0.29	+	P				
TTA	2.43	+++	P	L	8.91	++	P	CGC	0.46	+	V				
TTG	1.57	++	P					CGG	0.04	+	V				
CTA	0.51	++	V					CGT	0.63	++	V				
CTC	1.5	+++	V					CCA	2.62	++	V	P	4.42	+	N
CTG	0.34	+	V					CCC	0.18	+	V				
CTT	2.56	++	V					CCG	0.37	+	V				
								CCT	1.24	++	P				

The phases of periodicity are assigned based on the alignment of sequences from the translation start and define as peak (P), valley (V) and undistinguishable (N) based on the center position of the periodicity. In most of the cases, when a codon is ended with A or T, they are in the peak phase, and when the codon is ended with G or C, they are in the valley phase. Exceptions do exist for amino acids with three or more codons (shaded).

is transcript-centric, the length of 5′-UTR and introns are also interruptive. In *T. vaginalis* genome, there are only 65 introns identified so that the signal of the periodicity is rather maximal; it was also supported by the observation of a better signature after removal of intron-containing genes in *Entamoeba histolytica* (Supplementary Figure S4). This point is also footnoted by our failed attempts to discover significant periodicities in higher eukaryotes. Third, the better presentation of this periodicity among unicellular eukaryotes is believed to be largely due to the relatively regularity of nucleosome organization. For higher eukaryotes, the standard nucleosomes are composed of two copies of the four core histones H2A, H2B, H3 and H4 and the linker histone H1(12). H1 is critical in determining the high-order folding state of chromatins (42). There is a robust linear relationship between H1 stoichiometry and nucleosome repeat length, investigations in mammalian cerebral cortex demonstrated that neuronal and ganglial chromatins have nucleosome repeat length of 162 and 201 bp, and H1 stoichiometry of 0.45 and 1.04 molecule per nucleosome, respectively. This variation in nucleosome unit length might interrupt the formation of the periodicity. As the four core histones and nucleosome structure are highly conserved, histone H1 is rather variable and appears absent in some taxa (43). We identified five groups of histone genes in *T. vaginalis* through homolog search; the four core histones among eukaryotes were easily defined but the linker histone H1 was not easily defined. Similar results have been reported for *G. lamblia* (44,45) that also showed a well-defined sequence periodicity in this study.

Although we do not yet have direct evidence to illuminate how nucleosome organization gives rise to this periodicity, we can speculate a possible molecular mechanism. Histones and DNA helix are packaged into compact forms to prevent DNA from damages as opposed to the linker region between nucleosome folds, which is relatively exposed and thus vulnerable to mutagens. The susceptible linker DNA is damaged more frequently, repaired more often, and therefore leaves its nucleotide composition more variable. Since nucleosomes are not randomly positioned along DNA sequences and their regularity appears to coincide with transcriptional units, the different damage-repair frequency between linker-DNA and wrapping-DNA finally causes oscillation of nucleotide composition along nucleosome array over time. Nevertheless, this explanation requires further investigations but provides a plausible clue for formulating hypotheses and designing experiments.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr Weidong Bao for many constructive suggestions and our colleagues Jizeng Zhang, Shu Tao, Lei Wang and Jiucheng Liu for helpful discussions.

FUNDING

National Basic Research Program (973 Program) from the Ministry of Science and Technology of the People's Republic of China (2006CB910404 to J.Y.). Funding for open access charge: National Basic Research Program (2006CB910404).

Conflict of interest statement. None declared.

REFERENCES

1. Tuteja,N., Singh,M.B., Misra,M.K., Bhalla,P.L. and Tuteja,R. (2001) Molecular mechanisms of DNA damage and repair: progress in plants. *Crit. Rev. Biochem. Mol. Biol.*, **36**, 337–397.
2. Hu,J., Zhao,X. and Yu,J. (2007) Replication-associated purine asymmetry may contribute to strand-biased gene distribution. *Genomics*, **90**, 186–194.
3. Zhao,X., Zhang,Z., Yan,J. and Yu,J. (2007) GC content variability of eubacteria is governed by the pol III alpha subunit. *Biochem. Biophys. Res. Com.*, **356**, 20–25.
4. Worning,P., Jensen,L.J., Nelson,K.E., Brunak,S. and Ussery,D.W. (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima. *Nucleic Acids Res.*, **28**, 706–709.
5. Rhodes,D. and Klug,A. (1981) Sequence-dependent helical periodicity of DNA. *Nature*, **292**, 378–380.
6. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thastrom,A., Field,Y., Moore,I.K., Wang,J.P. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
7. Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
8. Watson,J.D. and Crick,F.H. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.
9. Noll,M. and Kornberg,R.D. (1977) Action of micrococcal nuclease on chromatin and the location of histone H1. *J. Mol. Biol.*, **109**, 393–404.
10. Rill,R.L., Oosterhof,D.K., Hozier,J.C. and Nelson,D.A. (1975) Heterogeneity of chromatin fragments produced by micrococcal nuclease action. *Nucleic Acids Res.*, **2**, 1525–1538.
11. Shaw,B.R., Herman,T.M., Kovacic,R.T., Beaudreau,G.S. and Van Holde,K.E. (1976) Analysis of subunit organization in chicken erythrocyte chromatin. *Proc. Natl Acad. Sci. USA*, **73**, 505–509.
12. McGhee,J.D., Felsenfeld,G. and Eisenberg,H. (1980) Nucleosome structure and conformational changes. *Biophys. J.*, **32**, 261–270.
13. Schalch,T., Duda,S., Sargent,D.F. and Richmond,T.J. (2005) X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, **436**, 138–141.
14. Herzel,H., Weiss,O. and Trifonov,E.N. (1998) Sequence periodicity in complete genomes of archaea suggests positive supercoiling. *J. Biomol. Struct. Dyn.*, **16**, 341–345.
15. Shepherd,J.C. (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl Acad. Sci. USA*, **78**, 1596–1600.
16. Tsonis,A.A., Elsner,J.B. and Tsonis,P.A. (1991) Periodicity in DNA coding sequences: implications in gene evolution. *J. Theor. Biol.*, **151**, 323–331.
17. Gutierrez,G., Oliver,J.L. and Marin,A. (1994) On the origin of the periodicity of three in protein coding DNA sequences. *J. Theor. Biol.*, **167**, 413–414.
18. Petrin,D., Delgaty,K., Bhatt,R. and Garber,G. (1998) Clinical and microbiological aspects of Trichomonas vaginalis. *Clin. Microbiol. Rev.*, **11**, 300–317.
19. Schwebke,J.R. and Burgess,D. (2004) Trichomoniasis. *Clin. Microbiol. Rev.*, **17**, 794–803.

20. Vanacova,S., Liston,D.R., Tachezy,J. and Johnson,P.J. (2003) Molecular biology of the amitochondriate parasites, Giardia intestinalis, Entamoeba histolytica and Trichomonas vaginalis. *Int. J. Parasitol.*, **33**, 235–255.
21. Carlton,J.M., Hirt,R.P., Silva,J.C., Delcher,A.L., Schatz,M., Zhao,Q., Wortman,J.R., Bidwell,S.L., Alsmark,U.C., Besteiro,S. *et al.* (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science*, **315**, 207–212.
22. Trifonov,E.N. (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A*, **249**, 511–516.
23. Yu,J. (2007) A content-centric organization of the genetic code. *Genomics Proteomics Bioinformatics*, **5**, 1–6.
24. Woodcock,C.L., Skoultchi,A.I. and Fan,Y. (2006) Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Res.*, **14**, 17–25.
25. Bailey,K.A., Chow,C.S. and Reeve,J.N. (1999) Histone stoichiometry and DNA circularization in archaeal nucleosomes. *Nucleic Acids Res.*, **27**, 532–536.
26. Musgrave,D., Forterre,P. and Slesarev,A. (2000) Negative constrained DNA supercoiling in archaeal nucleosomes. *Mol. Microbiol.*, **35**, 341–349.
27. Pereira,S.L., Grayling,R.A., Lurz,R. and Reeve,J.N. (1997) Archaeal nucleosomes. *Proc. Natl Acad. Sci. USA*, **94**, 12633–12637.
28. Sandman,K. and Reeve,J.N. (2000) Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. *Arch. Microbiol.*, **173**, 165–169.
29. Shivaswamy,S., Bhinge,A., Zhao,Y., Jones,S., Hirst,M. and Iyer,V.R. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.*, **6**, 618–630.
30. Wong,G.K., Wang,J., Tao,L., Tan,J., Zhang,J., Passey,D.A. and Yu,J. (2002) Compositional gradients in Gramineae genes. *Genome Res.*, **12**, 851–856.
31. Bulmer,M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature*, **325**, 728–730.
32. Knight,R.D., Freeland,S.J. and Landweber,L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.*, **2**, R10.
33. Stenico,M., Lloyd,A.T. and Sharp,P.M. (1994) Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acids Res.*, **22**, 2437–2446.
34. Xia,X. (1996) Maximizing transcription efficiency causes codon usage bias. *Genetics*, **144**, 1309–1320.
35. Fuglsang,A. (2004) The 'effective number of codons' revisited. *Biochem. Biophys. Res. Commun.*, **317**, 957–964.
36. Fuglsang,A. (2006) Estimating the "effective number of codons": the Wright way of determining codon homozygosity leads to superior estimates. *Genetics*, **172**, 1301–1307.
37. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
38. Bernstein,B.E., Liu,C.L., Humphrey,E.L., Perlstein,E.O. and Schreiber,S.L. (2004) Global nucleosome occupancy in yeast. *Genome Biol.*, **5**, R62.
39. Yuan,G.C., Liu,Y.J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, **309**, 626–630.
40. Liu,C.L., Kaplan,T., Kim,M., Buratowski,S., Schreiber,S.L., Friedman,N. and Rando,O.J. (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol.*, **3**, e328.
41. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
42. Hendzel,M.J., Lever,M.A., Crawford,E. and Th'ng,J.P. (2004) The C-terminal domain is the primary determinant of histone H1 binding to chromatin in vivo. *J. Biol. Chem.*, **279**, 20028–20034.
43. Kasinsky,H.E., Lewis,J.D., Dacks,J.B. and Ausio,J. (2001) Origin of H1 linker histones. *Faseb J.*, **15**, 34–42.
44. Wu,G., McArthur,A.G., Fiser,A., Sali,A., Sogin,M.L. and Mllerm,M. (2000) Core histones of the amitochondriate protist, Giardia lamblia. *Mol. Biol. Evol.*, **17**, 1156–1163.
45. Yee,J., Tang,A., Lau,W.L., Ritter,H., Delport,D., Page,M., Adam,R.D., Muller,M. and Wu,G. (2007) Core histone genes of Giardia intestinalis: genomic organization, promoter structure, and expression. *BMC Mol. Biol.*, **8**, 26.