Nuclear-Localized Plastid DNA Fragments in Protozoa, Metazoa and Fungi

Shu Yuan[a], Xin Sun[a], Lin-Chun Mu[b], Tao Lei[a], Wen-Juan Liu[a], Jian-Hui Wang[a], Jun-Bo Du[a], and Hong-Hui Lin[a,*]

[a] Key Laboratory of Bio-resources and Eco-environment (Ministry of Education), College of Life Science, Sichuan University, Chengdu, Sichuan 610064, P. R. China. Fax: 86-0 28-85 412571. E-mail: honghuilin@hotmail.com

[b] Department of Anatomy and Histo-embryology, Chengdu Medical College, Tianhui Road, Chengdu, Sichuan 610083, P. R. China

* Author for correspondence and reprint requests

Z. Naturforsch. **62c**, 123–132 (2007); received June 21/August 2, 2006

We analyzed nuclear-localized plastid-like DNA (nupDNA) fragments in protozoa, metazoa and fungi. Most eukaryotes that do not have plastids contain 40–5000 bp nupDNAs in their nuclear genomes. These nupDNA fragments are mainly derived from repeated regions of plastids and distribute through the whole genomes. A majority of nupDNA fragments is located on coding regions with very important functions. Similar to plastids, these nupDNAs most possibly originate from cyanobacteria. Analysis of them suggests that through millions of years of universal endosymbiosis and gene transfer they may have occurred in ancient protists before divergence of plants and animals/fungi, and some transferred fragments have been reserved till now even in modern mammals.

Key words: Nuclear-Localized Plastid-Like DNA (nupDNA), Endosymbiosis, Gene Transfer

Introduction

During endosymbiotic evolution, eukaryotic nuclear genomes have acquired numerous genes from the endosymbiotic organelles, which later evolved into the present chloroplasts and mitochondria (Kurland and Andersson, 2000; Martin *et al*., 2002). The eukaryotes that contain chloroplasts latterly evolved into plants, and others are called protozoa, metazoa and fungi. However, by recent discoveries, this discrimination is not necessarily the case. *Trypanosoma* and *Leishmania* parasites contain several plant-like genes encoding homologues of proteins found in either chloroplasts or the cytosol of plants and algae, pointing to a secondary loss of chloroplasts in trypanosomes (Martin and Borst, 2003; Hannaert *et al*., 2003). Two major apicomplexan parasites, *Plasmodium falciparum*, the infectious agent of malaria, and *Toxoplasma gondii*, which causes toxoplasmosis, were long known to contain an enigmatic organelle in their cytosol, called the hohlzylinder (apicoplast), which later was showed as a highly reduced chloroplast genome (McFadden *et al*., 1996). Recently, Okamoto and Inouye (2005) described a flagellate "Hatena", which acquires plastid by an endosymbiosis. However, this plastid was inherited by only one daughter cell. It is difficult to say whether it is an alga or a protozoon. All the phenomena imply that plastid symbiosis may more widely exist than we originally thought. Endosymbiosis should result in lateral gene transfer (Martin *et al*., 1998). "You are what you eat," wrote Doolittle (1998), when it comes to gene donations from organelles. However, how long will you be what you eat is still a question. If most protozoa once acquired some plastid sequences and subsequently evolved into metazoa (including mammals), a few nuclear-localized plastid-like DNA (nupDNA) fragments may be reserved even in higher animals. In this paper, we analyzes nupDNA fragments in protozoa, metazoa and fungi, suggesting that some plastid-originated sequences may be preserved in animals and fungi over 1000 million years (Myr) till now.

Methods

Complete sequences of *Oryza sativa* (X15901), *Marchantia polymorpha* (X04465) and *Porphyra purpurea* (U38804) chloroplast genomes were retrieved from GenBank. Using these sequences as query sequences, BLASTN searches were made for 42 eukaryotic genomes (Table I). Considering

0939–5075/2007/0100–0123 $ 06.00

that most of the sequences selected are derived form different species, *E*-values lower than 0.001 were defined as nupDNA fragments with biological meaning (Altschul *et al*., 1997). The redundancy of nupDNA fragments in overlapping regions of contigs was checked with information from the physical map of each species, if available. Short nucleotides repeat sequences were also filtered out. When estimating the total length of the nupDNA fragments in each species, the calculations were simplified by summing the lengths of the chloroplast genomic regions corresponding to individual nupDNA fragments. The number of intermingled nupDNAs was counted as the ones that contained discontinuous nupDNAs.

Then, nupDNA fragments of 40 eukaryotic genomes derived from BLAST comparisons for each chloroplast sequence were collected for secondary BLASTN with *Reclinomonas americana* mitochondrial genomes, *Oryza sativa* mitochondrial genomes, *Rickettsia felis* URRWXCal2, *Ehrlichia canis* str. *Jake*, *Wolbachia endosymbiont* of *Drosophila melanogaster*, *Nostoc* sp. PCC 7120, *Prochlorococcus marinus* subsp. *pastoris* str. CCMP1986, *Synechocystis* sp. PCC 6803 and other 60 bacterial genomes (*Aeropyrum pernix* K1, *Sulfolobus tokodaii* str. 7, *Pyrobaculum aerophilum* str. IM2, *Archaeoglobus fulgidus* DSM 4304, *Natronomonas pharaonis* DSM 2160, *Methanothermobacter thermautotrophicus* str. *Delta* H, *Methanococcus maripaludis* S2, *Methanopyrus kandleri* AV19, *Methanosarcina mazei* Go1, *Thermococcus kodakarensis* KOD1, *Thermoplasma volcanium* GSS1, *Nanoarchaeum equitans* Kin4-M, *Corynebacterium jeikeium* K411, *Mycobacterium tuberculosis* H37Rv, *Tropheryma whipplei* str. *Twist*, *Bacteroides thetaiotaomicron* VPI-5482, *Chlorobium tepidum* TLS, *Chlamydia trachomatis* D/UW-3/CX, *Chlamydophila pneumoniae* TW-183, *Bacillus cereus* E33L, *Listeria monocytogenes* str. 4b F2365, *Staphylococcus epidermidis* RP62A, *Clostridium tetani* E88, *Lactococcus lactis* subsp. *lactis* Il1403, *Streptococcus pyogenes* SSI-1, *Mycoplasma synoviae* 53, *Aquifex aeolicus* VF5, *Deinococcus radiodurans* R1, *Thermus thermophilus* HB8, *Magnetococcus* sp. MC-1, *Brucella suis* 1330, *Nitrobacter winogradskyi* Nb-255, *Rhodopseudomonas palustris* CGA009, *Sinorhizobium meliloti* 1021, *Anaplasma marginale* str. *St. Maries*, *Azoarcus* sp. EbN1, *Dechloromonas aromatica* RCB, *Nitrosospira multiformis* ATCC 25196, *Bordetella pertussis Tohama* I, *Burkholderia mallei* ATCC 23344, *Ralstonia solanacearum* GMI1000, *Neisseria meningitidis* Z2491, *Desulfovibrio desulfuricans* G20, *Geobacter sulfurreducens* PCA, *Campylobacter jejuni* RM1221, *Helicobacter pylori* J99, *Legionella pneumophila* str. *Paris*, *Psychrobacter arcticus* 273-4, *Thiomicrospira crunogena* XCL-2, *Buchnera aphidicola* str. *Sg*, *Escherichia coli* K12, *Salmonella typhimurium* LT2, *Yersinia pestis* KIM, *Haemophilus influenzae* 86-028NP, *Pseudomonas putida* KT2440, *Vibrio fischeri* ES114, *Xanthomonas oryzae* pv. *oryzae* KACC10331, *Xylella fastidiosa Temecula*1, *Borrelia garinii* Pbi, *Treponema pallidum* subsp. *pallidum* str. *Nichols*). For comparisons between chloroplast and mitochondrial genomes, 'Blast 2 Sequences' were used. The drop-off point of *E*-value was also 0.001 for this biological meaning (Altschul *et al*., 1997). The redundant sequences were filtered out. The results were simplified by summing the lengths of matches for each species.

Results

Plastid-like DNA fragments in eukaryotes without plastids

To evaluate the abundance of nupDNAs in eukaryotes who do not have plastids, we used *Oryza*, *Marchantia* and *Porphyra* chloroplast genomes as query sequences to search nupDNAs in 40 protozoa, metazoa and fungi nuclear genome databases, whose genome projects are almost complete at present. We identified >270 candidate sequences for each plastid BLASTN (Altschul *et al*., 1997) with biological meaning and *E*-values of 0.001. (Considering that most of sequences selected are derived form different species, *E*-values lower than 0.001 were used. Redundant candidate sequences were excluded, if physical map informations of contigs are available). The combined length of these fragments in each species is shown in Table I. Lengths of nupDNA fragments in eukaryotes without plastid usually ranged from 30 bp to 300 bp, and some *E*-values were less than 1×10^{-10}. None of the 40 selected eukaryotic genomes exceed 4000 Mb. DNA only contains the four bases A, G, C, T. Therefore, a maximum 16 bp ($\log_4 4 \times 10^9$) fragment for a certain sequence could be found only by chance. 30–300 bp nupDNA fragments with such high similarity acquired by BLAST researches should not be randomly found sequences. Furthermore, there is no common fragment that can be found in all the eu-

Table I. Combined length and number of nupDNA fragments in each eukaryotic genome. The table shows the total lengths of nupDNA fragments in each eukaryotic genome (number of nupDNA fragments).

BLAST subject nuclear genome sequences		BLAST query plastid sequences		
		Oryza sativa	*Marchantia polymorpha*	*Porphyra purpurea*
Plants	*Arabidopsis thaliana*	1.2×10^4 (117)	5034 (54)	3033 (31)
	Oryza sativa	9.1×10^5 (1632)	3.1×10^5 (1359)	9.2×10^4 (520)
Mammals	*Homo sapiens*	74 (1)	0 (0)	59 (1)
	Pan troglodytes[a]	74 (1)	0 (0)	59 (1)
	Bos taurus[a]	43 (1)	43 (1)	79 (1)
	Canis familiaris[a]	0 (0)	0 (0)	190 (3)
	Mus musculus	0 (0)	0 (0)	100 (2)
	Rattus norvegicus[a]	0 (0)	0 (0)	0 (0)
	Sus scrofa[a]	0 (0)	0 (0)	0 (0)
Animals	*Gallus gallus*[a]	0 (0)	65 (1)	130 (2)
	Danio rerio[a]	86 (2)	123 (3)	0 (0)
	Takifugu rubripes[a]	0 (0)	0 (0)	0 (0)
	Ciona intestinalis[a,b]	32 (1)	32 (1)	312 (5)
	Drosophila melanogaster	47 (1)	0 (0)	62 (1)
	Bombyx mori Dazao[a,b]	342 (7)	313 (7)	64 (1)
	Anopheles gambiae[a]	911 (11)	933 (10)	1588 (16)
	Aedes aegypti[a]	188 (4)	186 (4)	281 (5)
	Apis mellifera[a]	0 (0)	385 (3)	532 (7)
	Caenorhabditis elegans	0 (0)	0 (0)	74 (1)
	Caenorhabditis briggsae[a]	31 (1)	48 (1)	90 (1)
Fungi	*Aspergillus fumigatus*[a]	0 (0)	0 (0)	0 (0)
	Gibberella zeae[a]	0 (0)	0 (0)	0 (0)
	Neurospora crassa[a]	0 (0)	0 (0)	0 (0)
	Saccharomyces cerevisiae	178 (2)	166 (2)	857 (12)
	Yarrowia lipolytica	0 (0)	0 (0)	0 (0)
	Candida glabrata	0 (0)	0 (0)	235 (3)
	Schizosaccharomyces pombe	134 (1)	62 (1)	600 (7)
	Cryptococcus neoformans	80 (2)	80 (2)	0 (0)
	Encephalitozoon cuniculi	96 (3)	96 (3)	68 (2)
	Rhizopus oryzae[a,b]	371 (5)	1001 (13)	1517 (17)
Protists	*Leishmania major* str. *Friedlin*	252 (6)	252 (6)	252 (6)
	Trypanosoma cruzi[a,b]	556 (8)	492 (7)	516 (7)
	Trypanosoma brucei	210 (5)	210 (5)	446 (9)
	Theileria parva[a]	157 (3)	309 (4)	118 (3)
	Cryptosporidium parvum[a]	113 (2)	42 (1)	491 (9)
	Plasmodium falciparum	38 (1)	741 (6)	1002 (12)
	Plasmodium yoelii yoelii[a,b]	2459 (32)	3672 (42)	4171 (58)
	Dictyostelium discoideum[a]	56 (2)	87 (3)	571 (9)
	Entamoeba histolytica[a,b]	1277 (15)	1545 (13)	4694 (62)
	Giardia lamblia[a]	0 (0)	74 (1)	0 (0)
	Tetrahymena thermophila[a]	170 (1)	181 (2)	275 (4)
	Trichomonas vaginalis[a]	813 (9)	788 (9)	781 (8)

[a] The eukaryotic genome project is not complete at present.
[b] The length and the number of nupDNA fragments may be over-counted, because of lacking the physical map information of contigs.

karyotic genomes, which rule out the possibility that these candidate sequences are not real plastid fragments but have significant sequence similarities in all organisms. For example, a 90-bp ATP synthase α subunit fragment ($E = 1 \times 10^{-12}$ in *Caenorhabditis briggsae*) was found in only 10 of 18 mammal/animal nuclear genomes. Another 198-bp ATP synthase β subunit fragment ($E =$

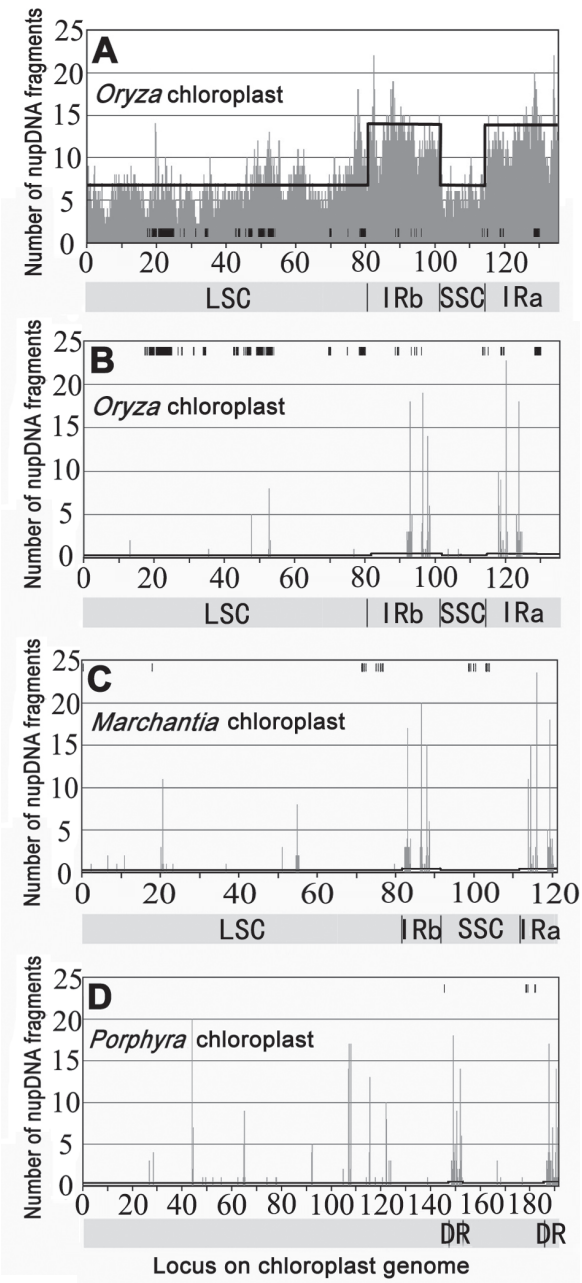

Fig. 1. Frequency of the appearance of nupDNA fragments throughout chloroplast genomes. The chloroplast genomes were divided into 100-bp segments. The numbers of nupDNA fragments corresponding to individual segments are shown by histograms. The *Oryza* or *Marchantia* chloroplast genome is double-stranded circular DNA, which contains two copies of an identical inverted repeat (IRa and IRb) separated by a large single-copy region (LSC) and a small single-copy region (SSC). The *Porphyra* chloroplast genome contains two copies of an identical direct rDNA repeat (DR). The black boxes indicate the regions whose copies are also found in their mitochondrial genome. The black line indicates the expected number of nupDNA fragments if they originated from throughout the chloroplast genome with equal frequency. (A) Frequency of the appearance of nuclear-localized *Oryza* plastid-like DNA fragments of the *Oryza* genome throughout the *Oryza* chloroplast genome. (B) Frequency of the appearance of nuclear-localized *Oryza* plastid-like DNA fragments of 40 eukaryotic genomes throughout the *Oryza* chloroplast genome. (C) Frequency of the appearance of nuclear-localized *Marchantia* plastid-like DNA fragments of 40 eukaryotic genomes throughout the *Marchantia* chloroplast genome. (D) Frequency of the appearance of nuclear-localized *Porphyra* plastid-like DNA fragments of 40 eukaryotic genomes throughout the *Porphyra* chloroplast genome.

8×10^{-22} in *Apis mellifera*) was found in *Apis mellifera* and *Anopheles gambiae*, but not in the other 16 mammal/animal nuclear genomes. 50- to 100-bp heat shock protein (HSP) fragments ($E = 8 \times 10^{-9}$ in *Saccharomyces cerevisiae*) was found in *Saccharomyces cerevisiae*, *Candida glabrata*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi*, but not in the other 5 fungal genomes. If these sequences are common fragments or selection-driven sequence convergences, then it is very difficult to explain why they distribute sporadically throughout the relative organisms. For most species compared with three plastids, most nupDNA fragments were found when *Porphyra* chloroplast genome was used as the query sequence. This is easy to explain, because *Porphyra* chloroplast genome is the biggest (191 kb) and most primitive one (Reith and Munholland, 1993). Another trend is that protozoa contain more nupDNAs than animals/fungi and mammals, suggesting that nupDNA fragments were continually lost during evolution. It is interesting that significantly long nupDNAs exist in *Anopheles* (a kind of mosquito). The combined length of nupDNA fragments in *Anopheles* is >1.5 kb, constituting 0.001% of the *Anopheles* genome. This could not happen only by chance.

Distribution of nupDNAs throughout plastid genomes and eukaryotic genomes

We investigated the distribution of nupDNA fragments on nuclear and original plastid genomes. For the rice genome (used as a control genome), nupDNA fragments originated from every part of the chloroplast genome at a similar frequency (Fig. 1A; on average, 6.7 times from single-copy regions), suggesting that transfers and inte-

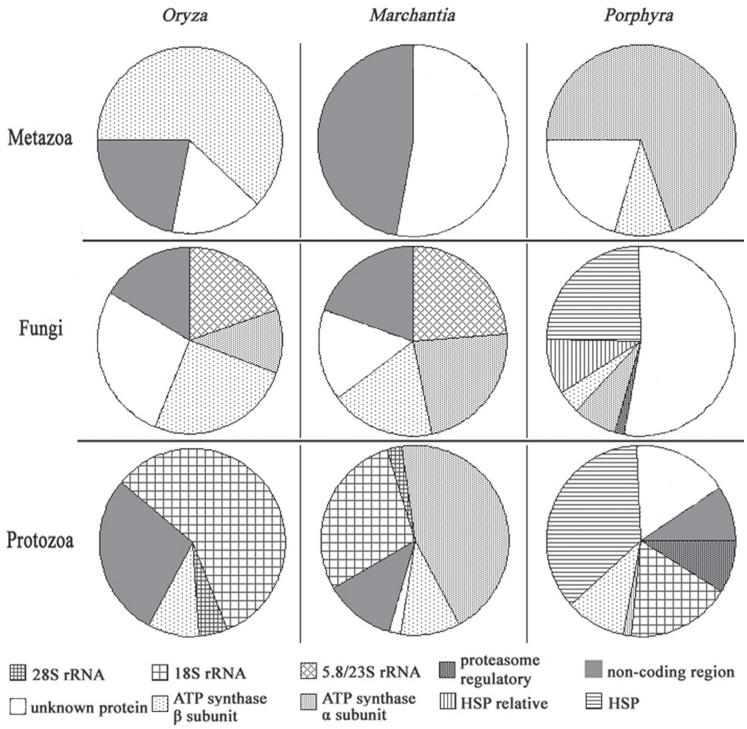

Fig. 2. Functional analysis of nupDNA fragments. NupDNA fragments with known functions in protozoa, metazoa and fungi derived from BLAST caparison with *Oryza*, *Marchantia* and *Porphyra* plastid genomes are collected in pie charts and marked with different hatchings. Each hatched sector indicates the combined length of nupDNA fragments with the same function.

grations into the nuclear genome occur almost equally throughout the rice chloroplast genome (Matsuo *et al.*, 2005). However, fewer nupDNA fragments can be found in 40 eukaryotic genomes, and the expected numbers of nupDNAs throughout each plastid are only about 0.2 (0.4 times from repeated regions), if they originated from each chloroplast with equal frequency (Figs. 1B–D). Although nupDNAs of 40 eukaryotic genomes are distributed throughout all parts of the plastid genome, they have some gene-transfer-hotspots, contrasting to gene transfers in the rice genome. Fragments in repeated regions (whether inverted repeat or direct repeat) prefer to transfer and integrate into the nuclear genome with higher frequencies. These regions usually encode rRNAs. About 20% of nupDNAs are rDNA fragments (see Fig. 2). The black boxes in Fig. 1 indicate the chloroplast DNA segments of which copies are found in the corresponding mitochondrial genome. For the rice plastid genome, the number of rice nupDNA fragments tends to be a little higher for those boxed regions (Fig. 1A). Notsu *et al.* (2002) suggested that mitochondrial genome might engulf plastid DNA and transfer it to the nucleus in flowering plants, which may frequently occur in rice. However, no such trend can be applied to nupDNAs in 40 eukaryotic genomes. On the contrary, for each mitochondrial genome of 40 selected eukaryotes, almost no similar nupDNA sequence can be found in each mitochondrion through BLAST search (data not shown). Thus, nupDNA fragments in 40 eukaryotic genomes may be transferred from the plastid and directly absorbed by the nucleus.

The integration sites on each genetic map (if available) were also investigated. NupDNA fragments are scattered throughout the chromosomes. We did not find any hotspot sites for whether long fragments or short fragments. This result is different to the findings in rice that large nupDNAs preferentially localize to the pericentromeric region of the chromosomes (Matsuo *et al.*, 2005). It also can be easily explained that after over 500

million years only a few nupDNA fragments have been reserved, and the distribution pattern was eliminated.

Possible origination of nupDNAs

To see the origination of these nupDNA fragments, we compared nupDNAs in protozoa, metazoa and fungi with 66 bacterial genomes. Considering that mitochondrion also originates from a prokaryotic organism and continually transfers genes to the nucleus (Adams *et al*., 2000) and there are a lot of genes common to mitochondria and plastids, it is also necessary to rule out the possibility that these nupDNAs are mitochondrion-originated sequences. The comparison results are shown in Fig. 3. Rice plastid genome was used as a control sequence to compare with the rice mitochondrial genome (a very large mitochondrial genome, 491 kb; Notsu *et al*., 2002), *Reclinomonas americana* mitochondrial genome (one of the most primitive mitochondrial genomes, 69 kb; Kurland and Andersson, 2000), three cyanobacterial genomes (putative ancestors of plastids; Martin *et al*., 2002), three Rickettsiales genomes (putative ancestors of mitochondria; Kurland and Andersson, 2000) and other 60 bacterial genomes. Except for the rice mitochondrial genome (also see Fig. 1A, the reason has been discussed above), rice plastid genome is most similar to cyanobacterial genomes, especially the genome of *Nostoc*, which is the most possible ancestor of the plastid (Martin *et al.*, 2002). For nupDNAs in protozoa and fungi, a similar pattern was seen and the longest homologues were found in cyanobacterial genomes. It is difficult to judge whose genome is more similar with metazoa nupDNAs. This may be due to the predilection of nupDNAs in metazoa. A large part of metazoa nupDNAs are ATP synthase fragments (Fig. 2), which may deviate BLAST comparison. Besides two mitochondrial genomes, nupDNAs in each species were also compared with their own mitochondrial genomes. Usually few homologues can be found in these comparisons (data not shown). In summary, nupDNA fragments in protozoa, metazoa and fungi have the same origination with modern chloroplasts. They should not originate from ancient mitochondrial genomes, not their own mitochondrial genomes, but cyanobacterial genomes. Through Fig. 1D as mentioned above, some fragments of genes unique to chloroplasts were also found, such as *trnG*, *trnT* and *tufA*, also suggesting that these nupDNAs more likely originate from ancient plastid/cyanobacterial genomes than mitochondrial genomes.

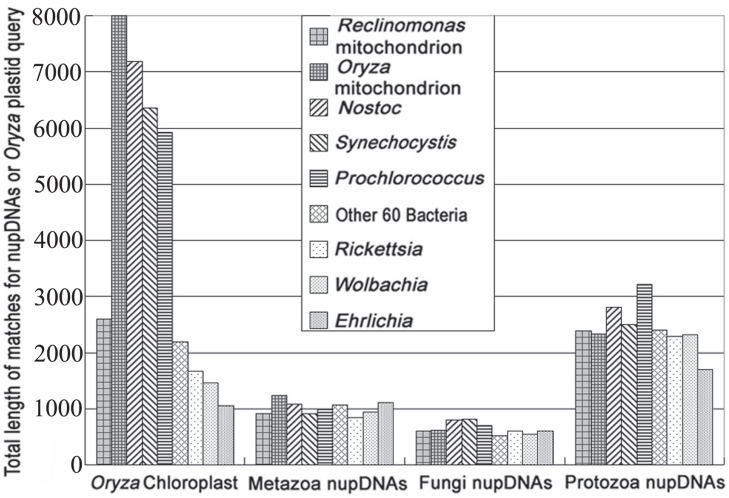

Fig. 3. Similarity of *Oryza* plastid genome sequence and nupDNA fragments in protozoa, metazoa and fungi to *Reclinomonas* and *Oryza* mitochondrial genomes, *Rickettsia*, *Ehrlichia*, *Wolbachia*, *Nostoc*, *Prochlorococcus*, *Synechocystis* genomes and other 60 bacterial genomes. The similarities were simplified by summing the lengths of matches when BLAST was used at *E* value threshold 0.001. Green columns indicate the longest combined lengths of nupDNA fragments in 60 selected bacterial genomes (see Methods).

Functional analysis of nupDNA fragments

Functional analysis demonstrates that most nupDNA fragments in the nucleus are located on coding regions with very important functions. Due to limited information of gene functions of known genomes, we only analyzed functions of nupDNA fragments in 17 eukaryotic genomes (Fig. 2). These sequences encode the ATP synthase α subunit, ATP synthase β subunit, 5.8/23S rRNA, 18S rRNA, 28S rRNA, proteasome regulatory, heat shock protein, translation elongation factor, dehydrogenase, RNA polymerase, histidine-tRNA ligase and other unknown proteins. Although there are also a few sequences located on non-coding regions, they are usually parts of promoters of important genes. Therefore, these sequences also may be of important uses. As a whole, most nupDNA fragments can be fallen into four kinds: ATP synthase, heat shock protein and related proteins, rRNA and other functional sequences.

The functions of homologues in cyanobacterial genomes, plant plastids and eukaryotic genomes are in consensus. For example, fragments of ATP synthase and elongation factor Tu in all genomes have the same functions. Homologues of heat shock protein 70 fragments function as molecular chaperone DnaK in cyanobacterial genomes. Cell division protein fragments in cyanobacterial genomes function as ATP-dependent Zn proteases in plant plastids and as proteasome regulatory in eukaryotic genomes. Eukaryote and prokaryote have different ribosomes and different rRNAs. Therefore, 16S rRNA fragments in plastid and cyanobacterial genomes are used as parts of 18S rRNA in eukaryotic genomes. Similarly, 23S rRNA fragments are used as 5.8/23S or 28S rRNA in eukaryotic genomes. Here, a conclusion can be drawn that most nupDNA fragments conserved in the eukaryotic nucleus keep their original functions or are endued with similar functions. This is necessary. If the chloroplast DNA sequence is not functional in the nucleus, the rate of nucleotide substitution rate should be $4.0 - 5.6 \times 10^{-9}$/site per year (Ramakrishna *et al.*, 2002; Matsuo *et al.*, 2005), and half-lives of nupDNAs should be 0.5 – 2.2 Myr (Matsuo *et al.*, 2005). Animals and plants diverged before 500 Myr (Kutschera and Niklas, 2004). Therefore, non-functional plastid sequences in the nucleus should be eliminated or randomly substituted that no fragment could be found through BLAST search. Although the average length of the similar sequence stretches is about 100 bp, too short to be a complete protein-coding gene, but enough long for a functional region in a protein/rRNA. Chloroplasts have a prokaryotic codon, while eukaryota code proteins in their own way. Hence, it is not possibly that a complete plastid gene can be transferred into the nucleus which maintains its function. Reasonably, plastid-special genes are less likely preserved in the nucleus after millions of years, such as genes for photosystems or Calvin cycle enzymes. Once plastid DNAs are integrated into the nuclear genome, they are rapidly fragmented and vigorously shuffled, and a vast majority of them are eliminated within several million years (Matsuo *et al.*, 2005). Only a few short fragments of common functions (such as rRNA, ATP synthase, HSP) can be reserved and reused in the nucleus through billions of years. Selection on nupDNAs may not be strong at the beginning, since most of them should not be functional. However, a few fragments have been selected and utilized from thousands of transferred sequences during subsequent million years of constant selection. Then during formation and evolution of fungi and metazoa, these nupDNA fragments were continually lost. But the losses were unparallel, and different fragments have been persevered in different species.

Discussion

It is salient that *E*-values of a large part of nupDNA fragments found in our research are bigger than 10^{-10}, and much of them are shorter than 100 bp. It cannot be affirmed that any piece of DNAs with an *E*-value less than 0.001 is a plastid-originated fragment. Also we cannot completely exclude the possibility that some short nupDNAs with low threshold are deep paralogues of the plastid genes that retain high similarity. We used such a low threshold for BLASTN because the candidate nupDNAs transferred into the nucleus may be reserved millions of years and changed too much. With the *E*-value of 0.001, we can find more possible plastid-originated fragments. However, it is obvious that there are a lot of long nupDNA fragments with very high similarities in metazoa and fungi, which are most possibly real plastid-originated sequences, for instance, the 65-bp fragment in *Gallus gallus* ($E = 1 \times 10^{-8}$), 294-bp fragment in *Anopheles gambiae* ($E = 2 \times 10^{-81}$), 198-bp fragment in *Apis mellifera* ($E = 8 \times 10^{-22}$), 92-bp fragment in *Saccharomyces cere-*

visiae ($E = 6 \times 10^{-15}$), 200-bp fragment in *Schizosaccharomyces pombe* ($E = 4 \times 10^{-17}$) and 288-bp fragment in *Rhizopus oryzae* ($E = 3 \times 10^{-33}$). Furthermore, we know so little about the concrete contours of early eukaryote evolution (Embley and Martin, 2006) that one cannot just casually dismiss the possibility that the ancestral eukaryote possessed a plastid as absurd or otherwise out of the question. Through analysis of nupDNAs in eukaryotes without plastids, we propose as hypothesis that millions years of universal endosymbiosis and gene transfer may have occurred in ancient protists before divergence of plants and animals/fungi.

Existent protists support universal endosymbiosis

Besides DNA sequence analysis, existent protists also support our hypothesis of anciently universal endosymbiosis. Plastid endosymbiosis is a common event once happened in all plant ancestors (McFadden, 2001). Previously non-photosynthetic protest engulfed and enslaved a cyanobacterium. This eukaryote then gave rise to the red, green and glaucophyte algae. Some protists also engulfed an existing eukaryotic alga involving a secondary endosymbiotic event. The dinoflagellates have undergone tertiary (engulfment of a secondary plastid) and even quaternary endosymbiosis (Bhattacharya *et al*., 2003). However, there are relatively few reports about endosymbiosis in protozoa. *Trypanosoma* and *Leishmania* were considered to have plastid in their evolutionary history (Martin and Borst, 2003). However, only 200- to 600-bp nupDNAs were found in their genomes (Table I). Giving that except for *Giardia lamblia* all other 11 protozoa contain plastid sequences longer than 200 bp (Table I), we estimate that probably over 90% of protozoa once had plastids through primary endosymbiosis or secondary endosymbiosis. Leander (2004) suggested that chloroplasts arose relatively recently within a specific subgroup of euglenids (relatives of trypanosomatid parasites). Okamoto and Inouye (2005) also demonstrated that a secondary symbiosis of green algae in a flagellate is in progress at present. These two instances indicate that plastid endosymbiosis is a common process in protists even under current natural conditions. Retrospecting to the Proterozoic era when eukaryotes emerged, universal endosymbiosis occurred. Heterotrophy may not prevail at that time (Kutschera and Niklas, 2004). Many protists adopted amphitrophy and temporarily contained some plastids, which may be like the flagellate Okamoto and Inouye observed. Hundred millions years later, some of the ancient protists evolved into protozoa, metazoa and fungi, and discarded plastids finally. However, there were also some protozoa reserved some plastid relicts, such as apicoplast in apicomplexa (Carlton *et al*., 2002; Abrahamsen *et al*., 2004; Gardner *et al*., 2005; Hall *et al*., 2005) and plate-like-chloroplast in *Ochromonas danica* (Semple, 1998). Besides, a few protists still kept their ability of engulfing photosynthetic eukaryotes heretofore, such as *Lotharella amoeboformis* (Ishida *et al*., 2000) and the flagellate "Hatena" (Okamoto and Inouye, 2005). From the first eukaryote naissance (1200 Myr ago; Butterfield, 2000) to the first metazoa appearance (570 Myr ago; Bengston, 1998), ancestors of metazoa and fungi should have much chance to acquire plastids. That is to say, ancestors of metazoa and fungi should have enough time to acquire plastid fragments. "You are what you eat," which means gene transfers from plastids, also happened in ancestors of metazoa and fungi. Now we come back the initial question "how long will you be what you eat?" Considering the nupDNAs in mammals, we estimate that it may be over 1000 Myr. Traditional view believes that plastid endosymbiosis only happened in ancestors of plants. But a lot of reports arising recently and our analysis of nupDNAs undermine this belief, and suggest that millions years of endosymbiosis and gene transfer occurred before the divergence of plants and animals/fungi.

Prospectively practical uses

It is notable that *Anopheles* and *Aedes* both are mosquitoes, however only *Anopheles* contains long nupDNA fragments. *Anopheles* is a vector of the *Plasmodium* that causes malaria (Holt *et al*., 2002). As mentioned above, *Plasmodium* is a protozoon that has a highly reduced plastid (McFadden *et al*., 1996). A plausible explanation is that *Anopheles* or an ancestor of *Anopheles* has a long-time contact with *Plasmodium*. During the course, apicoplast of *Plasmodium* transferred genes to nuclear then to *Anopheles*, or directly to *Anopheles*. Further investigation is needed to clarify this process. It is interesting that most eukaryotes who have long nupDNAs (>1 kb) are harmful parasites or their transmitting vectors. These insights led to the

discovery of some compounds that inhibit plant-specific pathways, much the way that herbicides do, also kill these parasites, and may suggest new targets for treating infections by these parasites (Fichera and Roos, 1997; Palenik, 2002). However, a mass of efforts about functional analysis of plant-like proteins/rRNAs still needs to be done before practical use of the information of eukaryotic nupDNAs. Only 40 eukaryotic genomes have been analyzed in this paper. We believe that more nupDNA fragments could be identified in the future accompanying more genome sequences available. Furthermore, how and when nupDNAs transferred to the nucleus and what happened to them after transfers still requires further research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (30571119) and the Program for New Century Excellent Talents in University of China. We thank Prof. Manyuan Long (University of Chicago, USA) for helpful discussion.

Abrahamsen M. S., Templeton T. J., Enomoto S., Abrahante J. E., Zhu G., Lancto C. A., Deng M., Liu C., Widmer G., Tzipori S., Buck G. A., Xu P., Bankier A. T., Dear P. H., Konfortov B. A., Spriggs H. F., Iyer L., Anantharaman V., Aravind L., and Kapur V. (2004), Complete genome sequence of the Apicomplexan, *Cryptosporidium parvum*. Science **304**, 441–445.

Adams K. L., Daley D. O., Qiu Y. L., Whelan J., and Palmer J. D. (2000), Repeat, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature **408**, 354–357.

Altschul S. F., Madden T. L., Schäffer A. A., Zhang J., Zhang Z., Miller W., and Lipman D. J. (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402.

Bengston S. (1998), Animal embryos in deep time. Nature **391**, 529–530.

Bhattacharya D., Yoon H. S., and Hackett J. D. (2003), Photosynthetic eukaryotes unite: endosymbiosis connects the dots. Bioessays **26**, 50–59.

Butterfield N. J. (2000), *Bangiomorpha pubescens* n. gen., n. sp: implications for the evolution of sex, multicellularity and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology **26**, 386–404.

Carlton J. M., Angiuoli S. V., Suh B. B., Kooij T. W., Pertea M., Silva J. C., Ermolaeva M. D., Allen J. E., Selengut J. D., Koo H. L., Peterson J. D, Pop M., Kosack D. S., Shumway M. F., Bidwell S. L., Shallom S. J., van Aken S. E., Riedmuller S. B., Feldblyum T. V., Cho J. K., Quackenbush J., Sedegah M., Shoaibi A., Cummings L. M., Florens L., Yates J. R., Raine J. D., Sinden R. E., Harris M. A., Cunningham D. A., Preiser P. R., Bergman L. W., Vaidya A. B., van Lin L. H., Janse C. J., Waters A. P., Smith H. O., White O. R., Salzberg S. L., Venter J. C., Fraser C. M., Hoffman S. L., Gardner M. J., and Carucci D. J. (2002), Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature **419**, 512–519.

Doolittle W. F. (1998), You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. **14**, 307–311.

Embley T. M. and Martin W. (2006), Eukaryotic evolution, changes and challenges. Nature **440**, 623–630.

Fichera M. E. and Roos D. S. (1997), A plastid organelle as a drug target in apicomplexan parasites. Nature **390**, 407–409.

Gardner M. J., Bishop R., Shah T., de Villiers E. P., Carlton J. M., Hall N., Ren Q., Paulsen I. T., Pain A., Berriman M., Wilson R. J., Sato S., Ralph S. A., Mann D. J., Xiong Z., Shallom S. J., Weidman J., Jiang L., Lynn J., Weaver B., Shoaibi A., Domingo A. R., Wasawo D., Crabtree J., Wortman J. R., Haas B., Angiuoli S. V., Creasy T. H., Lu C., Suh B., Silva J. C., Utterback T. R., Feldblyum T. V., Pertea M., Allen J., Nierman W. C., Taracha E. L., Salzberg S. L., White O. R., Fitzhugh H. A., Morzaria S., Venter J. C., Fraser C. M., and Nene V. (2005), Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. Science **309**, 134–137.

Hall N., Karras M., Raine J. D., Carlton J. M., Kooij T. W. A., Berriman M., Florens L., Janssen C. S., Pain A., Christophides G. K., James K., Rutherford K., Harris B., Harris D., Churcher C., Quail M. A., Ormond D., Doggett J., Trueman H. E., Mendoza J., Bidwell S. L., Rajandream M. A., Carucci D. J., Yates J. R. 3rd, Kafatos F. C., Janse C. J., Barrell B., Turner C. M., Waters A. P., and Sinden R. E. (2005), A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science **307**, 82–86.

Hannaert V., Saavedra E., Duffieux F., Szikora J. P., Rigden D. J., Michels P. A. M., and Opperdoes F. R. (2003), Plant-like traits associated with metabolism of *Trypanosoma* parasites. Proc. Natl. Acad. Sci. USA **100**, 1067–1071.

Holt R. A., Subramanian G. M., Halpern A., Sutton G. G., Charlab R., Nusskern D. R., Wincker P., Clark A. G., Ribeiro J. M., Wides R., Salzberg S. L., Loftus B., Yandell M., Majoros W. H., Rusch D. B., Lai Z., Kraft C. L., Abril J. F., Anthouard V., Arensburger P., Atkinson P. W., Baden H., de Berardinis V., Baldwin D., Benes V., Biedler J., Blass C., Bolanos R., Boscus D., Barnstead M., Cai S., Center A., Chaturverdi K.,

Christophides G. K., Chrystal M. A., Clamp M., Cravchik A., Curwen V., Dana A., Delcher A., Dew I., Evans C. A., Flanigan M., Grundschober-Freimoser A., Friedli L., Gu Z., Guan P., Guigo R., Hillenmeyer M. E., Hladun S. L., Hogan J. R., Hong Y. S., Hoover J., Jaillon O., Ke Z., Kodira C., Kokoza E., Koutsos A., Letunic I., Levitsky A., Liang Y., Lin J. J., Lobo N. F., Lopez J. R., Malek J. A., McIntosh T. C., Meister S., Miller J., Mobarry C., Mongin E., Murphy S. D., O'Brochta D. A., Pfannkoch C., Qi R., Regier M. A., Remington K., Shao H., Sharakhova M. V., Sitter C. D., Shetty J., Smith T. J., Strong R., Sun J., Thomasova D., Ton L. Q., Topalis P., Tu Z., Unger M. F., Walenz B., Wang A., Wang J., Wang M., Wang X., Woodford K. J., Wortman J. R., Wu M., Yao A., Zdobnov E. M., Zhang H., Zhao Q., Zhao S., Zhu S. C., Zhimulev I., Coluzzi M., della Torre A., Roth C. W., Louis C., Kalush F., Mural R. J., Myers E. W., Adams M. D., Smith H. O., Broder S., Gardner M. J., Fraser C. M., Birney E., Bork P., Brey P. T., Venter J. C., Weissenbach J., Kafatos F. C., Collins F. H., and Hoffman S. L. (2002), The genome sequence of the malaria mosquito *Anopheles gambiae*. Science **298**, 129–149.

Ishida K., Ishida N., and Hara Y. (2000), *Lotharella amoeboformis* sp. nov.: A new species of chlorarachniophytes from Japan. Phycol. Res. **48**, 221–230.

Kurland C. G. and Andersson G. E. (2000), Origin and evolution of the mitochondrial proteosome. Microbiol. Mol. Biol. Rev. **64**, 786–820.

Kutschera U. and Niklas K. J. (2004), The modern theory of biological evolution: an expanded synthesis. Naturwissenschaften **91**, 255–276.

Leander B. S. (2004), Did trypanosomatid parasites have photosynthetic ancestors? Trends Microbiol. **12**, 251–258.

Martin W. and Borst P. (2003), Secondary loss of chloroplast in trypanosomes. Proc. Natl. Acad. Sci. USA **100**, 765–767.

Martin W., Stoebe B., Goremykin V., Hansmann S., Hasegawa M., and Kowallik K. V. (1998), Gene transfer to the nucleus and the evolution of chloroplasts. Nature **393**, 162–165.

Martin W., Rujan T., Richly E., Hansen A., Cornelsen S., Lins T., Leister D., Stoebe B., Hasegawa M., and Penny D. (2002), Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. USA **99**, 12246–12251.

Matsuo M., Ito Y., Yamauchi R., and Obokata J. (2005), The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. Plant Cell **17**, 665–675.

McFadden G. I. (2001), Primary and secondary endosymbiosis and the origin of plastids. J. Phycol. **37**, 951–959.

McFadden G. I., Reith M. E., Munholland J., and Lang-Unnasch N. (1996), Plastid in human parasites. Nature **381**, 482.

Notsu Y., Masood S., Nishikawa T., Kubo N., Akiduki G., Nakazono M., Hirai A., and Kadowaki K. (2002), The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: Frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol. Genet. Gen. **268**, 434–445.

Okamoto N. and Inouye I. A. (2005), Secondary symbiosis in progress? Science **310**, 287.

Palenik B. (2002), The genomics of symbiosis: Hosts keep the baby and the bath water. Proc. Natl. Acad. Sci. USA **99**, 11996–11997.

Ramakrishna W., Dubcovsky J., Park Y. J., Busso C., Emberton J., SanMiguel P., and Bennetzen J. L. (2002), Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. Genetics **162**, 1389–1400.

Reith M. and Munholland J. (1993), A high-resolution gene map of the chloroplast genome of the red alga *Porphyra purpurea*. Plant Cell **5**, 465–475.

Semple K. T. (1998), Heterotrophic growth on phenolic mixtures by *Ochromonas danica*. Res. Microbiol. **149**, 65–72.