To: Prof. David Sankoff
Fax: 514-343-2254

From: Takashi Kunisawa
Science Univ. of Tokyo

total 6 pages

Dear David,
I hope you have time to go back
to the JME paper.
Takashi Kunisawa

VOL. 1, NOS. 2 & 3
1997

Research Communications in
Biochemistry and Cell & Molecular Biology

GENE ORDER COMPARISON FOR PHYLOGENETIC INFERENCE OF PLASTID GENOMES

Takashi Kunisawa[1]*, *Mathieu Blanchette*[2], *and David Sankoff*[2]

[1]Department of Applied Biological Sciences, Science University of Tokyo, Noda 278, Japan

[2]Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7, Canada

*Author for Correspondence and Reprints. E-mail: kunisawa@rs.noda.sut.ac.jp

Abstract

Gene order data of plastid genomes were compiled. To specify which gene is present and which is absent, homologous genes among plastid genomes were identified at the sequence level. For each pair of plastid genomes, the number of genome rearrangements, i.e., inversions, transpositions and inverted transpositions, was computed. It was shown that the ratio of rearrangements to the number of genes compared is the suitable measure for the phylogeny of land plant plastids. We also explored the contribution of the number of genes deleted (or inserted) to the evolution of algal plastid genomes

Introduction

Phylogenetic analysis of nucleic acid sequences traditionally compares homologous versions of a single gene in different organisms. This practice suffers from the arguable assumption that evolution of a single gene accurately represents evolution of the whole genome (Gray, 1988). As a complementary approach to resolve phylogenetic relationships, we have proposed a mathematical modeling of evolution at the genomic level (Sankoff, Leduc et al., 1992; Blanchette, Kunisawa et al., 1996). Phylogenetic distance between genomes is measured by gene order rearrangement based on the minimal set of inversions, transpositions, insertions and deletions necessary to convert the gene order in one genome to that of the other. The latter approach has been applied to a very limited case, i.e., mitochondrial genome evolution, simply because of the unavailability of complete gene-order data. In recent years, however, a number of plastid genomes are completely sequenced. These sequence data constitute another model for studying evolution due to genome rearrangement. Here we examine the validity and usefulness of the gene order analysis for the evolution of plastid genomes.

A nearly uniform number of genes (ca. 35) is found in animal mitochondrial genomes, and their phylogeny has been accounted for by the optimal number of moves, i.e., inversions, transpositions and inverted transpositions, that are responsible for the difference in gene order between genomes (Sankoff, Leduc et al., 1992). In contrast, the number of genes present in plastid genomes differs considerably; the plastid of nonphotosynthetic parasitic flowering plant *Epifagus virginiana*, in an extreme case, lack nearly 70 genes for photosynthesis and chlororespiration. In general the algal plastid genomes contain 1.5 to 2 times higher number of genes than land plant plastid genomes. We explore phylogenetic distance suitable for the evolutionary analysis of such genomes with nonuniform number of genes.

Data and Methods

Using the World Wide Web, gene order data were compiled from the existing major DNA databases GenBank/EMBL/DDBJ (for DDBJ, URL: http://www.ddbj.nig.ac.jp). As listed in Table I, currently 11 plastid genomes are completely sequenced. We have also compiled the gene order for a cyanobacterium *Synechocystis* sp. PC6803 (Kaneko, Sato et al., 1996), which was treated as an outgroup in constructing phylogenetic trees for plastids. Homologous genes were identified by the computer program FASTA (Lipman and Pearson, 1985). The minimal number of inversions, transpositions and inverted transpositions necessary to convert the gene order in one genome to that in the other was computed by the DERANGE2 program with the weight 2.05 of transpositions and inverted transpositions to inversions (Blanchette, Kunisawa et al., 1996). The PHYLIP package (Felsenstein, 1989) was also used in constructing phylogenetic trees.

Results and Discussion

Currently, complete nucleotide sequence data are available for five lineages of plastids, which show distinctive pigment compositions and ultrastructures (Table I). The first is a chlorophytic (green) lineage characterized by chlorophylls *a* and *b*, and stacked thylakoids. This lineage consists of the plastid from a green alga, *Chlorella vulgaris* (M. Sugiura, unpublished data) and of plastids from land plants including *Oryza sativa* (Hiratuka, Shimada et al., 1989), *Zea mays* (Maier, Neckermann et al., 1995), *Nicotiana tabacum* (Shinozaki, Ohme et al., 1986), *Pinus thunbergii* (Wakasugi, Tsudsuki et al., 1994), and *Marchantia polymorpha* (Ohyama, Fukuzawa et al., 1986) (see Table I). This lineage also includes a nonphotosynthetic plastid from *Epifagus virginiana* (Wolfe, Morden et al., 1992a), in which nearly 70 genes (or ORFs) are lost in evolution (see also Table II). The second lineage is a chloroplast from *Euglena gracilis* (Hallick, Hong et

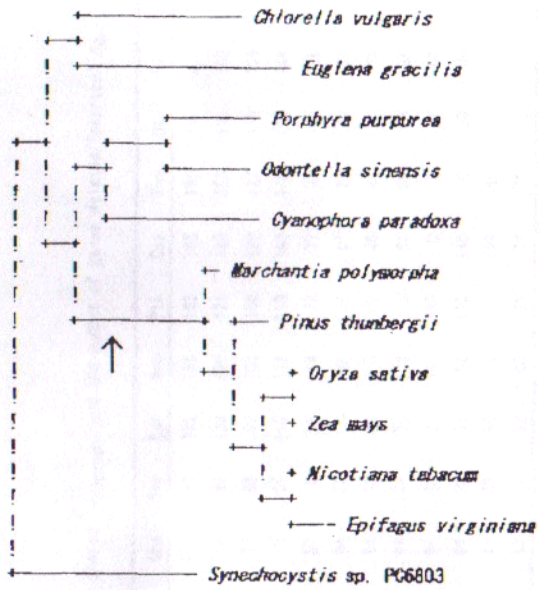

Fig. 1. Phylogenetic tree derived from the distance f_{move} between plastid genomes in Table III. The root is placed between a cyanobacterium, *Synechocystis* sp., and plastids. An arrow indicates the phylogenetic position of the *C. vulgaris*-*E. gracilis* clusterfor the case where $n_{deletion}$ distance is taken into account (see text).

Table I. Plastid genomes compared.

Genome	Accession Number[1]	Size (bp)	Genes (+ORFs)[2]
Chlorophytes			
green alga			
Chlorella valugaris	-	150,613	112
(M.Sugiura, unpublished)			
land plants			
Oryza sativa	X15091	134,525	157
Zea mays	X86563	140,387	157
Nicotiana tabacum	Z00044	155,854	135
Epifagus virginiana	M81884	70,028	55
Pinus thunbergii	D17510	119,707	116
Marchantia polymorpha	X04465	121,024	124
Euglenophyte			
Euglena gracilis	X70810	143,170	96
Chromophyte			
Odontella sinensis	Z67757	117,704	160
Rhodophyte			
Porphyra purpurea	U38804	191,028	193
Glaucophyte			
Cyanophora paradoxa	U30821	135.599	177

[1]GenBank/EMBL/DDBJ accession number

[2]Duplicated genes are counted twice and pseudogenes are not included

Table II. The number of genes shared, n_{shared}, and the number of genes deleted/inserted, n_{del}.

	Syn	Ppu	Osi	Cpa	Cvu	Egr	Mpo	Pth	Osa	Zma	Nti	Evi
Syn	0	47	120	111	154	182	155	182	183	185	166	215
Ppu	198	0	39	52	99	123	120	121	160	150	131	162
Osi	143	157	0	77	88	92	103	108	129	127	110	135
Cpa	156	158	130	0	95	115	112	113	140	140	123	148
Cvu	102	103	92	97	0	54	49	28	77	79	60	85
Egr	85	88	87	84	82	0	69	72	95	93	78	79
Mpo	107	98	90	94	93	90	0	29	40	42	21	70
Pth	104	98	88	94	104	79	109	0	55	55	38	85
Osa	107	97	91	94	93	81	117	110	0	8	27	84
Zma	108	97	92	94	92	82	118	110	148	0	23	88
Nti	107	98	92	94	93	82	118	110	129	131	0	85
Evi	54	54	51	53	52	52	65	58	72	71	73	0

n_{shared} is to the lower left and n_{del} is to the upper right. Syn, *Synechocystis* sp. PCC6803; Ppu, *Porphyra purpurea*; Osi, *Odontella sinensis*; Cpa, *Cyanophora paradoxa*; Cvu, *Chlorella vulgaris*; Egr, *Euglena gracilis*; Mpo, *Marchantia polymorpha*; Pth, *Pinus thunbergii*; Osa, *Oryza sativa*; Zma, *Zea mays*; Nti, *Nicotiana tabacum*; Epi, *Epifagus virginiana*

Table III. The number of recombinations, n_{move}, and its fraction f_{move} to n_{shared}.

	Syn	Ppu	Oei	Cpa	Cvu	Egr	Nbo	Pth	Oea	Zma	Nti	Evi
Syn	0	129	83	102	69	59	67	72	68	65	69	37
Ppu	0.658	0	51	77	63	61	64	58	54	54	64	32
Oei	0.580	0.325	0	71	54	61	47	62	50	49	46	29
Cpa	0.654	0.484	0.546	0	62	53	45	59	55	55	49	34
Cvu	0.675	0.612	0.587	0.639	0	53	56	68	54	52	51	34
Egr	0.684	0.693	0.701	0.631	0.646	0	50	49	53	53	52	41
Nbo	0.626	0.551	0.522	0.479	0.802	0.825	0	9	10	11	5	10
Pth	0.692	0.592	0.591	0.628	0.654	0.820	0.083	0	12	13	7	11
Oea	0.636	0.557	0.549	0.585	0.581	0.554	0.085	0.109	0	1	7	8
Zma	0.613	0.557	0.533	0.585	0.565	0.548	0.095	0.118	0.007	0	8	8
Nti	0.646	0.561	0.500	0.521	0.548	0.634	0.042	0.084	0.054	0.081	0	4
Evi	0.585	0.593	0.589	0.642	0.654	0.788	0.154	0.190	0.111	0.113	0.065	0

al., 1993), which is also characterized by chlorophylls *a* and *b* but is surrounded by three membranes. This feature of three envelope membranes has been interpreted as an indication of acquisition of a eukaryotic plastid (secondary endosymbiosis). The third is a plastid from a brown alga, *Odontella sinensis* (Kowallic, Stobe et al., 1995), which is marked by chlorophylls *a* and *c* and comprises four envelope membranes. Plastids from red algae (rhodoplasts) are the fourth lineage characterized by chlorophyll *a* and phycobilin, and by unstacked thylakoids. This lineage includes *Porphyra purpurea* (Reith and Munholland, 1995), for which the plastid genome is completely sequenced. The last fifth is a cyanelle from *Cyanophora paradoxa* (Stirewalt, Michalowski et al., 1995). This organelle shows chlorophyll *a* and phycobilin, and unstacked thylakoids but retains a peptidoglycan wall as do cyanobacteria.

To quantitatively compare the genome structures, we first identified homologous genes between plastids on the basis of sequence data not of gene labels. The number $n_{shared}(i,j)$ of homologous genes that are found both in genome *i* and *j* is listed in the lower left of Table II. Here, we have also compiled the gene order data of a cyanobacterium *Synechocystis* sp. PCC6803 (Kaneko, Sato et al., 1996), which is believed to share an evolutionary ancestor with the plastids and is recently completely sequenced. Among a total of 3217 genes in *Synechocystis*, 246 genes have homologous counterparts in the plastid genomes. For a substantial portion (196 genes), homologous genes can be found in the red algal plastid *P. purpurea*, whereas a much smaller number of homologous genes (roughly 100) are present in the plastids from land plants. The upper right of Table II lists the number $n_{deletion}(i,j)$ of genes that are not present in either one of the genomes *i* and *j*. In the comparison between rice and maize plastids, only six genes are deleted from one of the genomes and these monocot plants form a close evolutionary relative. In the comparison between rice and tobacco plastids, however, 27 genes (mostly unidentified ORFs) are not found in either one of the genomes. This higher value would come from the monocot-dicot comparison.

Removing genes absent from the either member of the genome pair, we have computed the optimal number n_{move} of moves (inversions, transpositions and inverted transpositions) necessary to convert the gene order in one genome into the other with the DERANGE2 program. The value of n_{move} is listed in the upper right of Table III and the lower left indicates the fraction of n_{move} to n_{shared}, i.e., $f_{move} = n_{move} / n_{shared}$. It is clear in Table III the land plant plastids show very similar gene arrangements (at most 13 moves in the comparison between the black pine *P. thunbergii* and maize *Z. mays*). In contrast, 51 to 63 moves are necessary to account for the differences between the nongreen plastids. The conversions between the nongreen and green plastids are

explained by roughly 50 recombinations. A phylogenetic tree based on the f_{move} distance is shown in Fig. 1. Here, the tree was constructed with the neighbor-joining method and was rooted by regarding *Synechocystis* as an outgroup.

Fig. 1 demonstrates that the branching order within land plants corresponds perfectly to the accepted evolutionary knowledge, with the successively deeper branchings for the angiosperms, the conifer and the moss. The plastid genome of nonphotosynthetic flowering plant *E. virginiana* shares the most recent ancestor with the common tobacco *N. tabacum*, being consistent with the sequence-level analysis (Wolfe, Morden et al., 1992b). Thus, the phylogeny within land plants can be accounted for by the rearrangement distance f_{move}. For the phylogeny of algal plastids, the present analysis shows a close relationship between the red alga *P. purpurea* and brown alga *O. sinensis*, suggesting the possibility of second endosymbiosis for the evolutionary origin of the brown alga (Gray, 1992). In the derived tree topology, a cluster of *C. vulgaris* and *E. gracilis* first diverge and then the divergence between nongreen plastids and land plastids occurs. Our preliminary analysis at the sequence level, however, shows that the divergence between nongreen and green plastids occurs first and then land plants diverge from the green alga. This inconsistency can be removed by taking account of the deletion distance $n_{deletion}$; the arrow in Fig. 1 illustrates the evolutionary position of the *C. vulgaris* and *E. gracilis* cluster for a tentative case where the distance is defined as $f_{move} + 0.5\, n_{deletion} / (n_{deletion} + n_{shared})$.

In conclusion, we have illustrated that the f_{move} distance is the suitable measure for the phylogeny of land plant plastids, but we need more empirical studies for the estimation of contribution from $n_{deletion}$ to the phylogeny of distantly related genomes.

Acknowledgments

The authors would like to thank Professor M. Sugiura for sending us the unpublished data. They also thank Professor A. Tsugita for helpful suggestions.

References

Blanchette M, Kunisawa T, Sankoff D (1996) Parametric genome rearrangement. Gene 172:GC 11-17

Felsenstein J (1989) PHYLIP-phylogeny inference package (version 3.2). Cladistics 5:164-166

Gray, M.W. (1988) Organelle origins and ribosomal RNA. Biochem. Cell Biol. 66:325-348

Gray MW (1992) The endosymbiont hypothesis revisited. Int Cytol Rev 141:233-357

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acid Res 21:3537-3544

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C, Meng B, Li Y, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes account for a major plastid inversion during the evolution of cereals. Mol Gen Genet 217:185-194

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., and Tabata, S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3:109-136

Kowallik KV, Stobe B, Schaffran I, Kroth-Pancic P, Freier U (1995) The chloroplast genome of a chlorophyll a+c containig alga, *Odontella sinensis*. Plant Mol Biol Reptr 13:336-342

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435-1441

Maier RM, Neckermann K, Igloi GI, Kössel H (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol 251:614-628

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S, Inokuchi H, Ozeki H (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322:572-574

Reith M, Munholland J (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol Biol Reptr 13:333-335

Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci USA 89:6575-6579

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki T, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of

the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043-2049

Stirewalt VL, Michalowski CB, Löffelhardt W, Bohnert HJ, Bryant DA (1995) Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. Plant Mol Biol Reptr 13:327-332

Wakasugi T, Tsudsuki J, Ito S, Nakashima K, Tsudsuki T, Sugiura M (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci USA 91:9794-9798

Wolfe KH, Morden CW, Palmer JD (1992a) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA 89: 10648-10652

Wolfe, K.H., Morden, C.W., Ems, S. and Palmer, J.D. (1992b) Rapid evolution of the plastid translational aparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J. Mol. Evol. 35:304-317

SEPARATION AND CHARACTERIZATION OF THE MATURED PROTEINS OF ADENOVIRUS TYPE 2 VIRION

*Takao Kawakami [1, 2], Akira Tsugita [1, *], Keiichi Hosokawa [3] and Yoshio Ueno [2]*

[1] Research Institute for Biological Sciences, Science University of Tokyo
Yamazaki, Noda, 278 Japan
[2] Department of Toxicology and Microbial Chemistry, Science University of Tokyo
Funagawara, Shinjuku-ku, Tokyo, 162 Japan
[3] Institute of Bio-Microbiology, Minowa, Taito-ku, Tokyo, 110 Japan

** Author for Correspondence and Reprints. FAX: +81 471 22 1544*

Abstract

Adenovirus-2 virion proteins were separated by two-dimensional polyacrylamide gel electrophoresis. One hundred and forty three protein spots were characterized by both molecular masses and isoelectric points. Proteins were analyzed for their amino-terminal sequences and twenty matured proteins were identified, including seven amino-terminal modified ones.

Introduction

Adenovirus type 2 (Ad-2) is classified to the non-oncogenic group of adenoviruses. The virion consists of an icosahedral shell, known as the capsid, and the DNA-protein core complex (Nermut, 1984). The capsid is composed of two kinds of capsomers, hexon and penton, arranged in a 5 : 3 : 2 cubic symmetry (Nermut, 1984). A fiber projects from each penton, and the additional several minor capsid proteins are found in close association with the hexons and pentons. The core complex contains a linear double-stranded DNA molecule with 35,937 base pairs (Roberts, O'Neill *et al.*, 1984), with two terminal proteins linked covalently to both 5'-ends (Rekosh, Russell, *et al.*, 1977), and three kinds of basic core proteins. The DNA encodes an endopeptidase that plays essential role for the virion maturation and infectivity (Weber, 1976).

Two-dimensional electrophoresis (2-DE), originally developed by O'Farrell (1975), has been introduced as one of the simplest methods of separating proteins directly from organisms. Bjellqvist, Ek *et al.* (1982) developed immobilized pH gradients (IPG)