

Biased frequency of short inverted-repeats in orchid chloroplast genomes

Daniel PRAT

Université Lyon 1, UMR 5023 LEHNA, 69622 Villeurbanne Cedex, France
daniel.prat@univ-lyon1.fr

Abstract – Genomes are structured by the sequences of nucleotides and their expression depend on various factors including chromatin condensation and methylation, repeated sequences, promotor structure and organization. Chloroplast and mitochondria genomes inherited from bacteria with their specific organization in circle shape and of small size making their analyze easier than that of nuclear genomes. Since the first complete sequence of chloroplast genome obtained for *Nicotiana tabacum*, many other chloroplasts genomes have been completely sequenced including in orchids. These genomes have been compared according their gene content and sequence but not really according to the distribution of short inverses repeats. In this study, we analyzed completely sequenced chloroplast genomes of 85 orchid species by counting the number of 2-, 4- and 6-bp short inverted repeats called palindromes. Their frequency is significantly lower than expected according to random association of nucleotides, this is particularly the situation for TA, CG, GC, ACGT, CGCG and TTAA while CCGG and especially GATC are significantly more frequent than expected. Species of Genus *Cypripedium* and *Epipogium* exhibited specific deviation from expectations, may be in *Epipogium* because of its small chloroplast genome size. Deviation of palindrome frequency within orchids fit only partly phylogeny. Palindrome sequences can be involved in interaction between DNA and proteins and consequently their frequency deviation can be related with gene expression, important cell process including relationships with foreign organisms.

INTRODUCTION

Chloroplast genomes are available for many plant species, including orchids. Mutations are supposed to be randomly distributed within genomes and shared point mutations are used to determine relationships among species and build phylogenies. Short inverted repeats consists of few DNA bases, like AACGTT able to form hairpins, whose DNA structure can be recognized and involved in expression control of genes. These sequences can be thus subjected to selection. The purpose of the present study is to determine if short inverted-repeats are randomly distributed in orchid chloroplast genomes.

MATERIALS AND METHODS

Analysis of 85 complete chloroplast sequences available in orchids:

- Counts for short-inverted repeats of 2, 4 and six DNA bases using seqinr package for R;
- Detection of biased distribution according to AT/GC content;
- Multivariate analyses of observed /expected frequency using FactoMineR package for R.

RESULTS

- Inverted-repeats: in average 9.7% less frequent than expected.
- Deviation in inverted-repeats distribution varies according to the species and sequence (Table 1).
- Most species have similar patterns excepted mainly *Epipogium* and *Rhizanthella* (Figure 1) which have a very reduced chloroplast genome size.
- Species cluster only partly according to phylogeny (Figure 2).

Table 1. Counts of some inverted-repeats in chloroplast genomes.

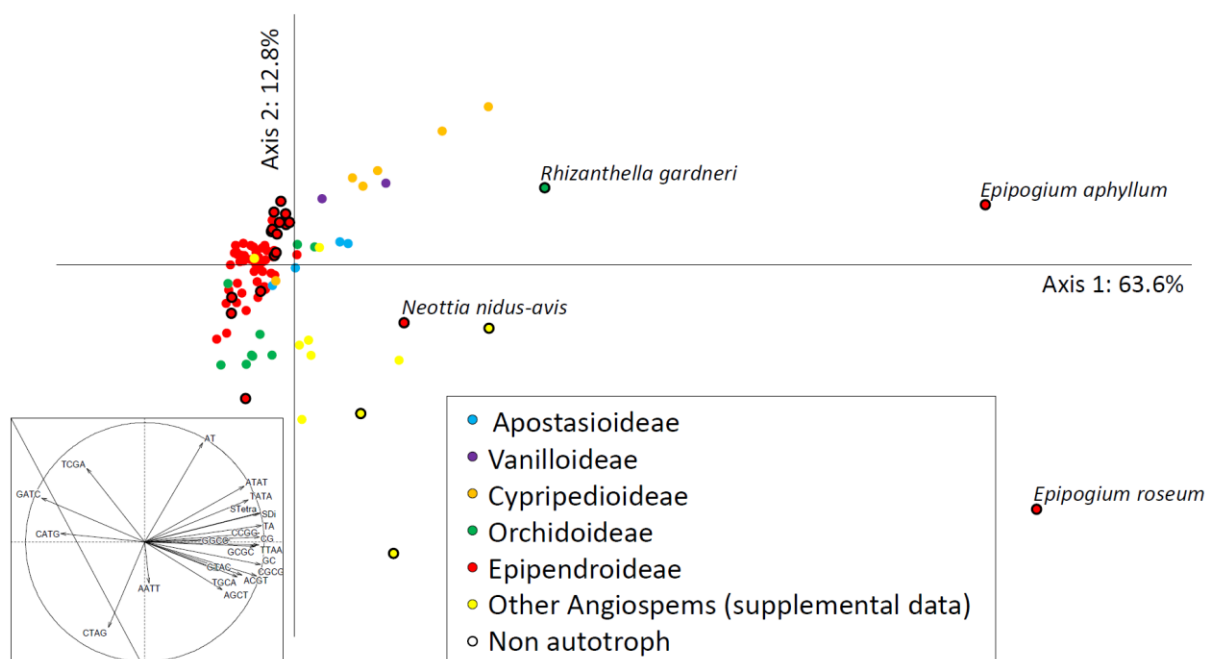
Taxon	2-bases	4-bases	6-bases	AT	TA	GC	ATAT	GATC	GCGC	GGCC	TATA	AACGTT	AGATCT	ATTAAT	TTTGAA
<i>Anoectochilus emeiensis</i>	36 177	9 395	2 542	15 407	12 050	4 406	1 469	792	121	237	1 143	171	67	88	82
<i>Anoectochilus roxburghii</i>	36 820	9 247	2 435	15 646	11 943	4 605	1 409	880	132	236	1 078	13	83	80	89
<i>Aphyllorchis montana</i>	22 562	5 736	1 589	9 668	7 395	2 700	928	515	86	141	725	93	53	51	54
<i>Apostasia odorata</i>	39 191	10 400	2 932	17 050	13 484	4 303	1 887	795	107	196	1 551	15	72	118	107
<i>Apostasia wallichii</i>	37 981	9 925	2 785	16 497	12 917	4 275	1 739	784	98	196	1 373	14	70	120	104
<i>Bletilla ochracea</i>	37 840	9 705	2 682	16 181	12 434	4 590	1 570	874	133	233	1 250	13	89	98	94
<i>Bletilla striata</i>	38 228	9 866	2 869	16 360	12 646	4 555	1 682	873	125	239	1 382	14	86	91	96
<i>Calanthe triplicata</i>	38 641	9 879	2 688	16 621	12 989	4 514	1 667	850	130	225	1 346	13	77	99	83
<i>Cattleya crispata</i>	35 350	8 884	2 398	15 187	11 621	4 253	1 451	831	127	217	1 110	12	76	82	81
<i>Cephalanthera humilis</i>	37 292	9 488	2 512	16 012	12 213	4 535	1 503	891	129	238	1 197	12	81	79	70
<i>Corallorhiza macrantha</i>	36 607	9 317	2 510	15 635	12 047	4 428	1 567	848	128	215	1 189	15	80	84	74
<i>Corallorhiza trifida</i>	36 149	9 187	2 475	15 436	11 917	4 388	1 548	849	123	219	1 186	13	85	90	80
<i>Cymbidium ensifolium</i>	35 914	9 236	2 564	15 457	11 848	4 277	1 536	828	120	212	1 207	14	77	103	79
<i>Cypripedium formosanum</i>	48 622	14 574	4 590	21 507	17 788	4 592	3 329	896	136	232	3 016	15	73	241	102
<i>Cypripedium macranthos</i>	37 807	9 691	2 572	15 856	12 283	4 655	1 451	886	131	249	1 109	19	82	107	100
<i>Dendrobium parviflorum</i>	35 463	8 937	2 414	15 080	11 517	4 360	1 426	842	126	227	1 088	13	85	83	88
<i>Elleanthus sodiroi</i>	38 556	9 801	2 653	16 540	12 767	4 546	1 595	897	126	225	1 225	14	86	99	100
<i>Epipactis veratrifolia</i>	37 999	9 649	2 607	16 275	12 451	4 618	1 540	870	130	243	1 176	12	69	89	85
<i>Epipogium aphyllum</i>	9 558	2 899	908	3 831	3 687	1 021	650	82	41	30	565	3	2	69	11
<i>Epipogium roseum</i>	5 862	1 836	631	2 345	2 419	575	406	37	17	18	414	3	4	32	7
<i>Erycina pusilla</i>	34 456	8 716	2 323	14 724	11 550	4 042	1 313	790	121	193	1 062	14	84	88	84
<i>Gastrochilus fuscopunctatus</i>	34 996	9 003	2 420	15 067	11 665	4 123	1 450	811	118	209	1 157	10	83	100	78
<i>Gastrochilus japonicus</i>	35 209	9 094	2 435	15 163	11 703	4 162	1 462	828	115	208	1 153	9	83	86	80
<i>Goodyera velutina</i>	36 108	9 313	2 523	15 428	11 951	4 423	1 489	797	122	237	1 133	14	65	93	76
<i>Habenaria pantlingiana</i>	37 668	9 806	2 704	16 104	12 660	4 485	1 595	821	125	230	1 346	12	75	99	88
<i>Liparis loeselii</i>	37 289	9 385	2 542	15 890	12 476	4 458	1 470	837	134	217	1 188	14	74	94	71
<i>Listera fugongensis</i>	36 893	9 353	2 534	15 718	12 023	4 602	1 431	881	141	242	1 102	13	82	86	78
<i>Ludisia discolor</i>	35 937	9 295	2 525	15 301	11 910	4 405	1 443	798	118	236	1 105	13	72	93	87
<i>Masdevallia coccinea</i>	38 086	9 579	2 560	16 314	12 785	4 489	1 540	875	130	222	1 171	13	85	100	81
<i>Neottia acuminata</i>	19 680	5 028	1 383	8 225	6 713	2 427	768	432	76	108	610	6	38	53	27
<i>Neottia camtschatea</i>	25 214	6 401	1 739	10 730	8 342	3 111	1 002	628	102	150	801	8	57	69	45
<i>Neottia nidus-avis</i>	23 861	6 485	1 850	10 393	8 798	2 447	1 240	458	53	111	1 119	5	40	81	26
<i>Neottia ovata</i>	36 885	9 310	2 495	15 717	11 987	4 611	1 392	877	134	242	1 072	12	76	87	89
<i>Neuwiedia singapureana</i>	40 434	11 013	3 285	17 526	13 988	4 447	2 024	820	113	214	1 840	13	68	136	96
<i>Oberonia japonica</i>	34 072	8 433	2 230	14 484	11 291	4 154	1 306	798	131	198	974	14	74	92	79
<i>Oncidium sphacelatum</i>	35 380	9 044	2 419	15 162	11 672	4 218	1 448	822	121	204	1 147	13	81	97	92
<i>Paphiopedilum niveum</i>	40 428	11 095	3 276	17 573	14 045	4 386	2 106	857	119	224	1 837	13	68	147	81
<i>Pelatanthera scolopendrifolia</i>	35 433	9 293	2 563	15 347	11 881	4 104	1 577	826	113	216	1 250	11	78	112	80
<i>Phalaenopsis aphrodite</i>	35 654	9 262	2 497	15 368	11 920	4 191	1 495	826	114	222	1 178	10	82	95	78
<i>Phragmipedium longifolium</i>	38 405	10 454	2 955	16 518	13 267	4 254	1 912	817	127	208	1 623	11	65	115	88
<i>Rhizanthella gardneri</i>	16 218	4 406	1 317	7 005	5 907	1 648	834	317	70	79	764	11	35	52	33
<i>Sobralia callosa</i>	38 751	9 871	2 716	16 638	12 867	4 548	1 617	901	129	226	1 284	13	90	90	103
<i>Thrixspermum japonicum</i>	36 544	9 682	2 702	15 852	12 465	4 116	1 736	810	114	208	1 383	10	82	116	77
<i>Vanilla aphylla</i>	38 553	10 466	3 042	16 835	13 484	3 955	2 108	741	104	206	1 832	13	79	132	112
<i>Vanilla planifolia</i>	36 941	9 964	2 866	16 274	12 517	3 911	1 912	755	109	194	1 543	15	83	111	125

 $\alpha = 0.0001$ $\alpha = 0.001$ $\alpha = 0.01$

Observed counts > expected counts

 $\alpha = 0.0001$ $\alpha = 0.001$ $\alpha = 0.01$

Observed counts < expected counts

**Figure 1.** Principal component analysis applied on inverted-repeats of chloroplast genomes.

CONCLUSION

Some inverted-repeat sequences are significantly more frequent than expected while some other ones are less frequent in chloroplast genome, independently from phylogeny. Their frequency appeared thus under selection pressure and possibly related to adaptation. *Epipogium* and *Rhizanthella*, underground orchids, have a peculiar pattern of inverted-repeats in complement to their

reduced chloroplast genome size. Deviation from expected distribution of inverted-repeats shows a signature for plant habitus according to their autotrophy pattern at adult stage. It may be related to their ability to contract relationships with other organisms (may be not only mycorrhizae).

Such selective effects on short DNA sequences could introduce deviation in phylogeny building (reduced probability to get TA, CG and CG succession in DNA bases).



Figure 2. Hierarchical cluster analysis applied on inverted-repeats of chloroplast genomes.