Analysis of proline codon usage in HRGPs of *Chlamydomonas reinhardtii*.

	number of codons-a	%GC -b	%GC if normal bias -c	%GC 3rd position -d	number of proline codons	% prolines	%CCG/C	%CCG	%CCC	%CCA	%CCT
Chlamy average	127,008	65	65	86	7392	6	80	31	49	8	12
SAD1 shaft	873	74	83	56	526	60	51	26	25	20	29
SAD1 N-term	612	62	64	82	42	7	74	36	38	19	7
SAD1 head	2404	72	74	78	221	9	70	36	34	13	17
SAG1 shaft	934	80	84	70	560	60	66	33	33	10	24
SAG1 N-term	469	63	65	79	37	8	67	32	35	27	5
SAG1 head	2006	73	74	80	167	8	81	54	27	11	8
GP1 shaft	342	79	84	66	208	61	64	16	48	3	34
GP1 rest	213	65	65	88	14	7	79	0	79	0	21
MTA2 shaft	166	78	83	66	100	60	62	27	35	17	21
MTA2 rest	220	61	65	75	12	6	59	17	42	7	25
GAS28 shaft	61	79	85	64	39	64	61	10	51	0	38
GAS28 rest	385	63	64	83	22	6	78	14	64	9	14
VSP1 shaft	190	70	73	77	61	32	70	26	44	8	21
VSP1 rest	110	66	70	74	4	4	50e	25	25	25	25

	number of codons-a	%GC -b	%GC if normal bias -c	%GC 3rd position -d	number of proline codons	% prolines	%CCG/C	%CCG	%CCC	%CCA	%CCT
VSP3 shaft	203	69	76	64	90	44	65	17	48	1	34
VSP3 rest	270	63	64	83	9	3	78	11	67	0	22
VSP4 shaft	66	80	78	89	34	52	88	32	56	6	6
VSP4 rest	925	65	64	89	58	6	94	34	60	3	2
ZSP1 shaft	123	73	74	81	42	34	81	29	52	5	14
ZSP1 rest	79	69	72	79	1	1	100e	100	0	0	0
GP2 shaft	241	79	85	65	167	69	60	19	41	0	40
GP2 rest	1017	65	65	86	85	8	79	14	65	1	20

Abbreviations:
a) number of codons in the indicated gene region
b) GC content of the mRNA segment(s) coding for the indicated gene region
c) the calculated GC content for the mRNA if the codons for all amino acids had been used in the same proportions as in an average *Chlamydomonas* gene
d) % of codons that have a G or a C in the third position
e) too few prolines for the result to be meaningful

Source of data:
Chlamy average: A compilation of *Chlamydomonas reinhardtii* genes obtained from a codon usage database maintained by the Kazusa DNA Research Institute, http://www.kazusa.or.jp
VSP1 L16461
VSP3 L29029
VSP4 AY036106
ZSP1 S44199
GP2 AY596305
GAS28 AF015883
GP1 AF309494
MTA2 AF309495
SAG1 AY450930
SAD1 AY450929

Further notes:
The codon usage has been determined separately for the proline-rich portion (termed “shaft”) of each protein (which combines several segments in some proteins) and for the remainder of the protein (termed “rest”). For the two agglutinin proteins (SAG1 and SAD1) the head domain and the N-terminal domain were calculated separately.

Codon usage in *Chlamydomonas reinhardtii* is biased, and favors codons ending in G or C over those ending in A or T (86% of the time), consistent with the overall GC richness of coding regions (on average 65.4% GC). In regions rich in proline codons (CCN), as are found in HRGPs, this could lead to extremely high GC contents. For example, a region of 100 prolines would, on average, contain only 8 CCA and 12 CCT codons, resulting in a GC content of 280/300 or 93%. The data document that codon bias in the proline rich regions of cell wall and agglutinin genes is attenuated, with far more CCA and CCT codons used than expected. For example, the *minus* agglutinin shaft, which is 60% proline codons, has a GC content of 74%. This is, however, less than the GC content would be if the codon usage were typical of an average *C. reinhardtii* gene, in which case the GC content would have been 83% (these calculations are done for the codons of all the amino acids in the given region, not just the prolines). Thus the frequency of GC in the third position of all codons is only 56% rather than the usual 86%, and only 51% of proline codons end in G/C, rather than the usual 80%. This effect is not necessarily restricted to proline codons: for example, alanine codons (GCN), which are also GC rich, end in G or C 77% of the time on average, but 24% of the time in the SAD1 shaft (not shown).

The reduced 3rd position GC content is seen in all the proteins with long proline-rich shafts (i.e. SAD1, SAG1, GP1, MTA2, and GP2; all > 60% prolines). The shaft region of GAS28 is shorter but quite proline-rich, and also shows the effect.

The effect is less consistent in shafts of lower proline content. The 3rd position GC content for proline in VSP4, which has a short shaft of alternating serine and proline, is actually above average. In ZSP1 the effect is not seen perhaps because the shaft is fairly short. There is some effect in VSP1 and VSP3, whose shafts, while only 30-40% proline, are of greater length than those of ZSP1 and VSP4.