CODON UTILIZATION, DNA LANDSCAPING AND FRACTAL ANALYSIS IN BACTERIOPHAGE Φadh

N.R. McEWAN

Rowett Research Institute, Greenburn Road, Bucksburn, Aberdeen, AB21 9SB Scotland

Received November 3, 2004; accepted June 14, 2005

Summary. – The bacteriophage Φadh has a low G+C content and encodes its protein products using a restricted number of the codons, which it could theoretically use. Investigated were (i) the restricted codon usage by determining codon indices and codon distances for various genes and ORFs, (ii) distribution of purines and pyrimidines on the two strands of the double-stranded genome and within all genes and ORFs, and (iii) nucleotide positional bias within the genome. The genes and ORFs can be clustered into four groups, based on codon distance analysis. The genome landscape showed that the plus strand was more purine-rich than the negative one and that the only area of the genome where the landscape was located in the pyrimidine-rich region was at the start of the sequence which was also the only region of the genome where ORFs were found on the negative strand. The nucleotide composition of the genome, examined by fractal analysis showed little, if any, DNA positional bias, as opposed to overall compositional bias with a self-similarity profile. The ORFs showed a bias in favour of purines on the coding strand.

Key words: bacteriophage Φadh; codon usage; DNA landscape; fractal analysis

Introduction

The benefits of using lactobacilli as probiotics have been known for almost a century (Metchnikoff, 1907) and continue to be an area of intensive research. Recent examples of the probiotic potential include the production of bacteriocins by a number of strains of *Lactobacillus gasseri*, which can inhibit growth of food-borne pathogens (Itoh *et al*., 1995; Tahara *et al*., 1997). In order that the potential of the lactobacilli may be maximized, there has been a recent focus on the bacteriophage that infects members of this genus. Two principal reasons exist for investigating the biology of these bacteriophage: (i) to understand the mechanism whereby the phage may prevent growth of a newly introduced and desirable strain; and (ii) to construct a vector for genetic studies by using a phage derived from an agent which infects the cells naturally.

One such bacteriophage is Φadh, which was first identified as a prophage in the *L. gasseri* ADH genome (Raya *et al*., 1989). It was described as a linear, double-stranded DNA of approximately 43 kbp (Raya *et al*., 1992). Complete sequences for a number of genes of this bacteriophage have been published (e.g. Fremaux *et al*., 1993; Henrich *et al*., 1999; Engel *et al*., 1998). More recently, a complete sequence of the genome of this bacteriophage has been determined (Altermann *et al*., 1999). The genome is 43,785 bp long, with 3' protruding ends of 12 nucleotides, G+C content of 35.6%, and a potential to encode up to 62 putative ORFs.

When the deduced amino acid sequences of these ORFs were compared, e.g. by BLAST or FASTA searches, with those deposited in databases earlier, a number of ORFs showed some similarity to the genes or ORFs of other genomes, generally the genomes of other bacteriophage (Table 1). In many cases the level of similarity was confined to a relatively small fragment of the putative Φadh ORF, meaning that a functional role could not be reliably ascribed

E-mail: n.mcewan@rowett.ac.uk; fax: +441224-716687.

Abbreviations: GC3 = use of G or C in position 3 of a codon; MADCA-BORU = mean absolute distance codon analysis based on residue utilisation; MRI = mutational response index; Nc = effective codon number

Table 1. Characteristics of genes (ORFs) of bacteriophage Φadh

Gene (ORF)	Start	End	Size (kb)	G+C	GC3	GC-GC3	MRI	Nc	Land-scape
Integrase (intG)	1246	89	385	0.307	0.220	0.087	0.192	44.7	↙
ORF B	1357	1590	77	0.333	0.231	0.103	—	—	↗
ORF A	2288	1404	294	0.314	0.190	0.124	0.254	40.3	↙
ORF 2	2809	2441	122	0.339	0.309	0.030	0.064	45.9	↙
Transcription repressor	3161	2835	108	0.361	0.266	0.095	0.173	60.0	↙
Tec	3598	3813	71	0.338	0.306	0.032	0.102	61	↙
ORF 159	3916	4395	159	0.369	0.288	0.081	0.086	43.2	↙
ORF 72	4398	4616	72	0.329	0.260	0.068	—	—	↙
ORF 70	4588	4800	70	0.305	0.254	0.052	—	—	↙
ORF 82	5045	4797	82	0.337	0.229	0.108	—	—	↙
ORF 127	5549	5166	127	0.286	0.195	0.091	0.199	44.0	↗
ORF 65a	5786	5983	65	0.323	0.242	0.081	—	—	↓
ORF 71b	5996	6211	71	0.361	0.306	0.056	—	—	↙
ORF 49	6221	6370	49	0.293	0.3	-0.007	—	—	→
ORF 55	6380	6547	55	0.321	0.339	-0.018	—	—	↙
ORF 116	6650	7000	116	0.365	0.333	0.031	0.010	32.9	↙
ORF 88	6993	7259	88	0.368	0.312	0.055	—	—	↙
DNA replication	7259	7930	223	0.313	0.368	-0.055	0.079	44.4	↙
ORF 188	7927	8493	188	0.351	0.233	0.118	0.183	41.3	↙
Hel}{case	8477	9844	455	0.349	0.259	0.091	0.174	43.0	↙
ORF 175	9991	10518	175	0.348	0.290	0.059	0.198	41.2	↙
ORF 77	10529	10762	77	0.338	0.282	0.056	—	—	↖
Primase	10776	13091	771	0.355	0.297	0.058	0.108	45.6	↙
ORF 68	13470	13676	68	0.329	0.348	-0.019	—	—	↙
Repressor protein	13673	14011	112	0.322	0.204	0.118	0.248	40.3	↙
ORF 208	14015	14641	208	0.332	0.234	0.097	0.180	43.9	↙
ORF 146	14654	15094	146	0.324	0.265	0.059	0.135	41.5	↙
ORF 114a	15091	15435	114	0.299	0.191	0.107	0.219	43.8	↙
ORF 65b	15462	15659	65	0.323	0.273	0.051	—	—	↙
ORF 118	15656	16012	118	0.403	0.311	0.092	0.094	39.5	↙
ORF 52	16045	16203	52	0.245	0.189	0.057	—	—	↘
ORF 73	16227	16448	73	0.288	0.230	0.059	—	—	↙
ORF 197	16517	17110	197	0.315	0.278	0.037	0.179	43.6	↙
ORF 126a	16968	17348	126	0.323	0.331	-0.008	0.017	61	↘
ORF 90	17332	17604	90	0.381	0.363	0.018	0.086	40.8	↙
ORF 163	17699	18190	163	0.366	0.256	0.110	0.132	45.9	↙
Muramidase, export control	18281	18748	155	0.361	0.288	0.073	0.164	47.9	↙
ORF 170	19066	19578	170	0.439	0.404	0.035	-0.015	56.6	↗
ORF 149, terminase small	19737	20186	149	0.376	0.313	0.062	0.078	45.1	↙
ORF 624, terminase large	20183	22057	624	0.338	0.246	0.091	0.162	45.1	↙
ORF 397, portal protein	22245	23438	397	0.376	0.234	0.142	0.158	44.8	↙
Clp protein	23389	24117	242	0.376	0.300	0.075	0.104	50.3	↙
Head protein	24118	25305	395	0.365	0.250	0.115	0.218	36.3	↙
ORF 126b	25322	25702	126	0.373	0.244	0.129	0.176	41.7	↙
ORF 126c	25659	26036	125	0.401	0.353	0.048	0.056	53.6	↙
ORF 159b	26008	26487	159	0.392	0.344	0.048	0.029	55.7	↙
ORF 123a	26471	26842	123	0.320	0.194	0.126	0.158	46.4	↙
Tail protein	26842	27555	237	0.391	0.269	0.122	0.119	48.6	↙
ORF 183	27570	28121	183	0.346	0.201	0.145	0.185	36.8	↙
ORF 87	28069	28332	87	0.402	0.318	0.083	—	—	↙
ORF 302, tail protein	28332	29240	302	0.378	0.224	0.154	0.130	41.7	↙
Capsid protein	29327	33790	1487	0.371	0.261	0.109	0.169	40.4	↙
ORF 247	33777	34520	247	0.348	0.242	0.106	0.176	44.7	↙
ORF 241	34517	35242	241	0.364	0.240	0.124	0.188	44.0	↙
ORF 731	35121	37316	731	0.393	0.292	0.101	0.119	46.4	↙
ORF 938 – Capsid protein	37472	40288	938	0.368	0.299	0.069	0.112	47.5	↙
ORF 154	40348	40812	154	0.406	0.329	0.077	0.104	47.2	↙
ORF 1	40825	41040	71	0.380	0.278	0.102	—	—	↙
ORF 123b	40943	41314	123	0.384	0.419	-0.035	-0.005	61	↖
ORF 69	41394	41603	69	0.362	0.286	0.076	—	—	↙
Hol	41600	41944	114	0.328	0.200	0.128	0.172	39.2	↙
Lysin	41948	42901	317	0.398	0.258	0.141	0.128	43.1	↙
ORF C	43444	43629	61	0.296	0.177	0.118	—	—	↙

Start, End = nucleotide position; GC3 = use of GC in position 3 of the codon; GC-GC3 = difference between GC and GC3 values (i.e. GC3 subtracted from GC); landscapes occupying the bottom left quadrant have the symbol "↙"; those in the top right quadrant have the symbol "↗"; those which follow the Y-axis downward have the symbol "↓"; those which increase along the X-axis have the symbol "→". For other abbreviations see their list on the front page.

to the sequence. Thus, in the absence of functional analysis of the different putative ORFs, it was impossible to determine if these sequences encode a functional protein or not. All that can be deduced about this sequence at present is that the majority of the putative ORFs have the potential to encode a functional protein. Therefore the following analysis was performed, unless otherwise stated, based entirely on the assumption that any ORF was only hypothetical.

Materials and Methods

Downloading of bioinformatical resources. The complete bacteriophage Φadh genome sequence (Acc. No. AJ131519) was downloaded from the NCBI Public Database (http://www.ncbi.nlm.nih.gov). All ORFs were defined as those identified within this database for this bacteriophage. The codon usage table for *L. gasseri* genomic sequences was downloaded from the Japanese Codon Usage Database (http://www.kazusa.or.jp/codon).

Sequence similarity searches. The similarity between the derived protein sequences of the various ORFs and those already described was determined by BLASTP analysis using the EBI Public Database (http://www.ebi.ac.uk).

Pyrimidine-purine walking, genome landscaping and gene landscaping. A "pyrimidine-purine walk" (Lobry, 1999) was performed along the length of the genome. This was done by starting on the X-axis and simultaneously moving along the X-axis for every nucleotide in the genome at the same time as moving 1 unit upwards on the Y-axis for every *C* or *T* nucleotide and 1 unit downwards on the Y-axis for every *A* or *G* nucleotide. A genome landscape (Lobry, 1999) was constructed by starting the plot at the origin of the graph, and movements were made according to the nucleotide as follows: move 1 unit North for every *T* nucleotide, 1 unit South for every *A* nucleotide, 1 unit West for every *G* nucleotide and 1 unit East for every *C* nucleotide. A method similar to genome landscaping, designated gene landscaping was performed using the same technique, only restricting the landscape analysis to an individual gene. In all cases, gene landscapes were performed only on the coding strand of the DNA.

Effective codon number (Nc) and mutational response index (MRI) Two codon indices, Nc (Wright, 1990) and MRI (Gatherer and McEwan, 1997) were calculated for all ORFs potentially encoding proteins of 90 or more residues (amino acids) in length.

Mean absolute distance codon analysis based on residue utilization (MADCA-BORU). The frequency of usage for each codon capable of encoding a residue (amino acid) was calculated. Three types of average were used: (i) that for ORFs which showed similarity to genes of a known function, (ii) that for ORFs which showed during BLAST analysis similarity to other ORFs in the database, and (iii) that for all ORFs in the genome. The expected frequency of each codon was calculated on the basis of the number of each residue encoded in each ORF and the absolute value of the difference between this number and the observed number calculated. The values were summed for each amino acid, and the total for that amino acid was calculated. The various totals for all amino acids were then summed to give a MADCA-BORU value. This method deviates from those used in previous studies where either Euclidean or Hamming distances (e.g. Garcia-Vallve *et al*., 2000) were used. Whereas previous studies were focused on the genes with similar functions and thereby using similar amino acids, recent comparative studies across a complete genome implicate that many pair-wise comparisons between ORFs require comparing sequences, which are not similar in their amino acid content. For this reason MADCA-BORU was used to take account of codon usage based on the amino acid composition. The differences between the MADCA-BORU values for ORFs were calculated in a pairwise manner and formulated in grid format. The resulting grid was used as the input file for the NEIGHBOR program within the PHYLIP suite of programs (Felsenstein, 1989). The tree file generated was viewed using the TreeView (Page, 1996). MADCA-BORU trees were constructed for each of the three types of mean values (see above), each with a different base-line or standard reference point.

Fractal analysis of the genome. A fractal plot of the genome was performed using Microsoft Excel. The top right of the graph was designated "T" (1.1), bottom right "G" (1.0), bottom left "A" (0.0), and top left "C" (0.1). The plot started in the centre of the graph (0.5, 0.5). The next point on the graph was plotted as halfway between this point and the corner corresponding to the first nucleotide (e.g. if the nucleotide was *G*, then the new plot would be at (0.75, 0.25). Plots were then made halfway between the new plot and the corresponding corner, until the complete genome had been analysed.

Results and Discussion

Sequence similarity

When the deduced amino acid sequences of the genes (ORFs) of bacteriophage Φadh were compared (using e.g. BLAST or FASTA searches) with the protein sequences available in databases, a number of them showed some similarity to genes or ORFs from other genomes, generally from the genomes of other bacteriophage (Table 1). In many cases the level of similarity was confined to a relatively small fragment of the putative Φadh ORF, meaning that a functional role could not be reliably ascribed to the sequence. Thus, in the absence of functional analysis of the different putative ORFs it is impossible to decide if these sequences encode a functional protein or not. All that can be deduced about this sequence at present is that the majority of the putative ORFs have the potential to encode a functional protein. Therefore the analysis of codon usage in the bacteriophage Φadh genome was performed.

Codon usage

This analysis, unless otherwise stated, was based entirely on the assumption that any ORF was only hypothetical. It can be predicted for a genome with a G+C content deviating markedly from 50% that certain codons are used preferentially

Table 2. Codon usage patterns for all ORFs in bacteriophage Φadh and host genomes

Residue	Codon	Phage	Cell	Difference	Residue	Codon	Phage	Cell	Difference
Phe	TTT	0.729	0.753	–0.024	Ser	TCT	0.298	0.249	0.049
	TTC	0.271	0.247	0.024		TCC	0.039	0.026	0.013
Tyr	TAT	0.706	0.535	0.172		TCA	0.249	0.404	–0.155
	TAC	0.294	0.465	–0.172		TCG	0.052	0.038	0.014
Cys	TGT	0.652	0.593	0.059		AGT	0.229	0.183	0.046
	TGC	0.348	0.407	–0.059		AGC	0.134	0.100	0.034
His	CAT	0.665	0.537	0.128	Leu	TTA	0.426	0.444	–0.017
	CAC	0.335	0.463	–0.128		TTG	0.167	0.206	–0.040
Asn	AAT	0.742	0.699	0.043		CTT	0.202	0.205	–0.002
	AAC	0.258	0.301	–0.043		CTC	0.037	0.028	0.009
Asp	GAT	0.783	0.695	0.088		CTA	0.126	0.076	0.050
	GAC	0.217	0.305	–0.088		CTG	0.042	0.041	0.000
Gln	CAA	0.788	0.938	–0.150	Arg	CGT	0.246	0.389	–0.143
	CAG	0.212	0.062	0.150		CGC	0.061	0.147	–0.086
Lys	AAA	0.643	0.506	0.137		CGA	0.130	0.049	0.081
	AAG	0.357	0.494	–0.137		CGG	0.027	0.049	–0.022
Glu	GAA	0.778	0.888	–0.110		AGA	0.456	0.325	0.131
	GAG	0.222	0.112	0.110		AGG	0.080	0.042	0.038
Pro	CCT	0.392	0.379	0.013	Ile	ATT	0.649	0.657	–0.007
	CCC	0.074	0.041	0.033		ATC	0.161	0.188	–0.027
	CCA	0.442	0.543	–0.101		ATA	0.189	0.156	0.034
	CCG	0.092	0.037	0.055					
Thr	ACT	0.486	0.690	–0.204	Trpe	TGG	1.000	1.000	0.000
	ACC	0.110	0.082	0.027	Met	ATG	1.000	1.000	0.000
	ACA	0.301	0.172	0.129					
	ACG	0.104	0.056	0.048	Stops	TAA	0.561	0.650	–0.089
Val	GTT	0.420	0.504	–0.084		TAG	0.136	0.300	–0.164
	GTC	0.083	0.049	0.034		TGA	0.303	0.050	0.253
	GTA	0.352	0.381	–0.029					
	GTG	0.145	0.066	0.079					
Ala	GCT	0.488	0.498	–0.010					
	GCC	0.087	0.081	0.005					
	GCA	0.374	0.389	–0.015					
	GCG	0.052	0.032	0.019					
Gly	GGT	0.400	0.475	–0.075					
	GGC	0.201	0.208	–0.007					
	GGA	0.319	0.270	0.049					
	GGG	0.079	0.047	0.032					

in encoding proteins, as determined by the abundance of either GC or AT in position 3 of codons. As expected, the general pattern of codon usage was similar to that determined for the few sequences in the host organism (*L. gasseri*) so far published (Table 2). Only one codon (ACT) showed greater than 20% change in its abundance; in the bacteriophage, the usage of this codon in encoding threonine was 69%. It is also interesting to note that the usage of the stop codon TGA in the bacteriophage was much more common relative to the host, based on bacterial sequences described so far. However, over 50% of the codons showed less than a 5% change in their relative abundance, suggesting that the genome of the bacteriophage has evolved to have codons which are most likely to be recognized by the tRNAs of the host. In the case of this genome, the bias was generally in favour of A and T in position 3 (Table 1, column 4). However, there are 6 cases where the GC3 value was actually greater than that of the genome (ORF49, ORF55, DNA replication gene, ORF68, ORF126a, and ORF123b); these included at least one case where the ORF encoded a functional gene. In these cases the codons being used were not those, which were most abundantly used in this genome. Normally, this would be interpreted as an ORF, which is not likely to be functional. However, this could also be due to the gene requiring some form of retarded translational speed, e.g. by requiring unusual tRNAs, and thereby ensuring that proper protein folding takes place during translation.

MRI and Nc indices

In addition, two codon indices, MRI and Nc generally give patterns typical of those seen in gene prediction studies –

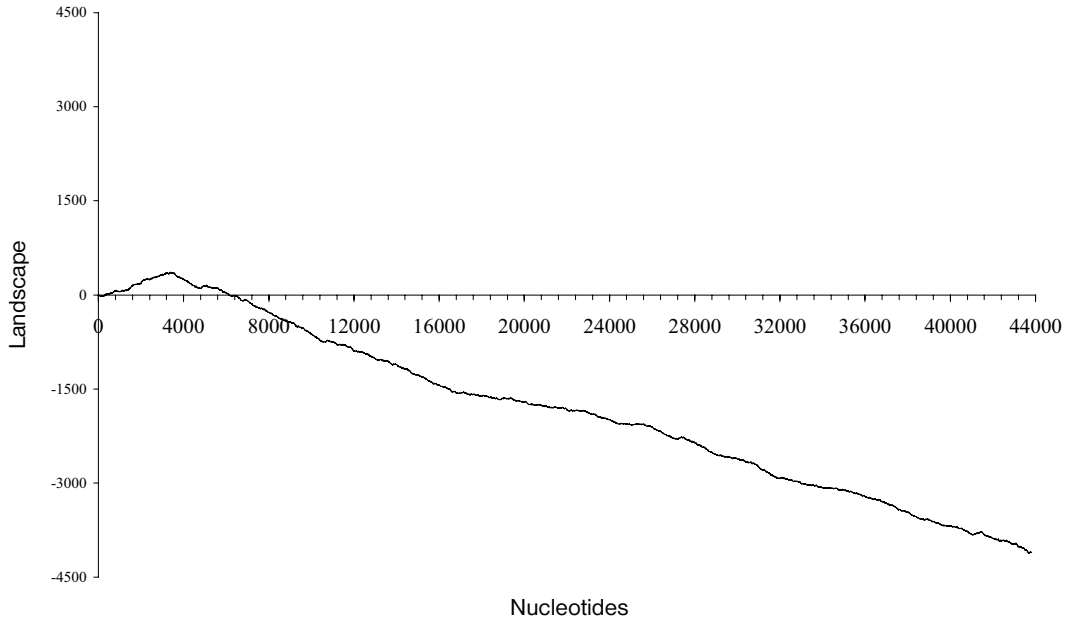

Fig. 1

The purine-pyrimidine walk along bacteriophage Φadh genome

Each pyrimidine resulted in a movement upwards on the Y-axis and each purine in a downward movement on this axis.

with low Nc values (maximum 61, minimum 20) and high (greater than 0.15) MRI values (McEwan and Gatherer, 1998, 1999), re-enforcing the suggestion that the majority of these theoretical ORFs are likely to encode genes. Examples of negative MRI values have been described previously (McEwan and Gatherer, 1999). It is assumed that these may, like the example discussed above for unusual GC3 usage, be seen as an indication of requiring unusual codons – possibly to ensure correct folding during translation. MRI and Nc are only two examples of a number of available indices. Others include the optimal codon-anticodon energy (Gouy and Gautier, 1982), intrinsic codon deviation index (Freire-Picos *et al*., 1994), and relative synonymous codon usage value (Sharp and Li, 1987). However, in previous work (McEwan and Gatherer, 1998, 1999) MRI and Nc were found to be the most useful of these indices in predicting ORF functionality for genes, which had a strong bias from the 50% G+C content.

Purine-pyrimidine walk

An alternative mechanism for studying the structure of the bacteriophage genome is using the purine-pyrimidine walk (Lobry, 1999). The pattern for the pyrimidine-purine walk of this phage is shown in Fig. 1. It is obvious that the positive strand is initially pyrimidine-rich, but it soon becomes soon purine-rich. Interestingly, the area where the trace lies above the X-axis (the start of the genome) is that where ORFs are present on the negative strand.

Genome landscaping

The technique of genome landscaping has proved useful as a tool for studying genomic DNA sequences and gave an indication of a genome duplication event in the evolutionary history of the spirochete *Borrelia burgdorferi*. The genome landscape of the bacteriophage Φadh is shown in Fig. 2. It is obvious that the landscape only occupies two quadrants of the graph; the upper right and the lower left, with the part of the landscape that is most remote from the origin lying in the lower left quadrant. This genome landscape pattern demonstrated that it was not the result of a bias either for or against a single nucleotide, but rather it was the product solely of a purine-pyrimidine bias.

A plot similar to that in Fig. 2, a gene landscape can be drawn for each individual ORF. This is plotted in relation to the direction of the putative start codon, but not relative to the ORF being on either the positive or negative strand. Over 80% of the ORFs showed a gene landscape similar to that seen for the complete genome (Table 1, column 8). Those deviating from this pattern were three gene landscapes occupying the lower right quadrant (ORF-52, ORF-126a and ORF-170); two gene landscapes primarily occupying the upper right quadrant (ORF-B and ORF-127), two gene landscapes primarily occupying the upper left quadrant (ORF-77 and ORF-123b), ORF-49 lying on the X-axis pointing West; and ORF-65a lying on the Y-axis pointing South. Of these nine putative ORFs only ORF-

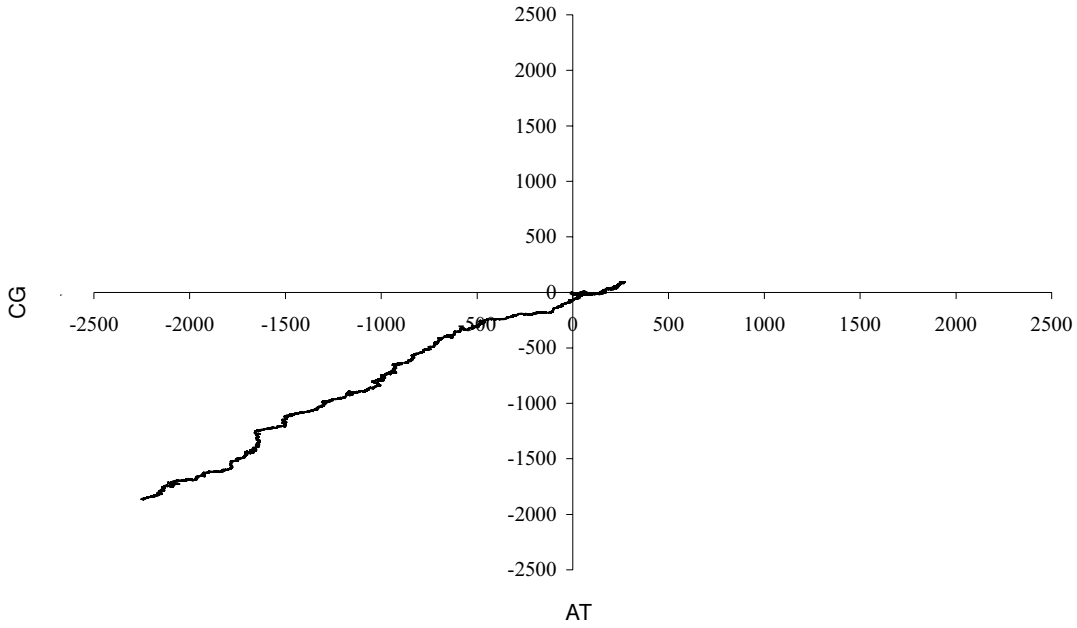

Fig. 2

The landscape of bacteriophage Φadh genome

The landscape was plotted by moving 1 unit North for every *T* nucleotide, 1 unit South for every *A* nucleotide, 1 unit West for every *G* nucleotide, and 1 unit East for every *C* nucleotide.

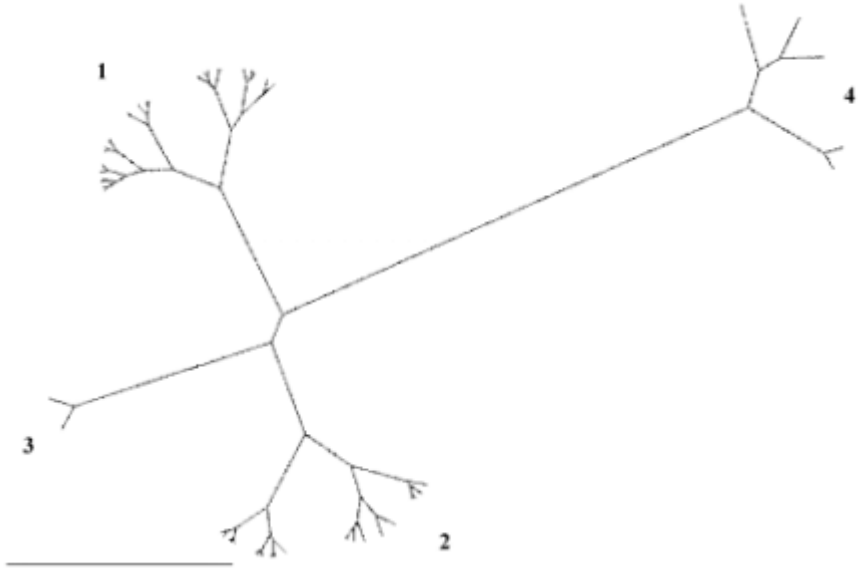

Fig. 3

The relationship between ORFs in bacteriophage Φadh genome as assessed by the use of MADCA-BORU values and genes as reference values

Group 1 (ORF 2, tec, ORF 72, ORF 70, ORF 82, ORF 127, ORF 65a, ORF 71b, ORF 116, ORF 188, ORF 146, ORF 65b, ORF 118, ORF 73, ORF 126a, ORF 90, ORF 155, ORF 170, ORF 126c, ORF 123a, ORF 87, ORF 154, ORF 1, ORF 123b, ORF 69 and ORF C), Group 2 (intG, ORF A, rad, ORF 159, ORF 88, ORF 223, ORF 455, ORF 175, ORF 112, ORF 208, ORF 114a, ORF 197, ORF 163, ORF 149, ORF 624, ORF 397, ORF 242, mhp, ORF 126b, ORF 159b, mtp, ORF 183, ORF 302, ORF 247, ORF 241, ORF 731, ORF 938, hol and lys), Group 3 (primase and capsid protein) and Group 4 (ORF B, ORF 49, ORF 55, ORF 77 and ORF 52) represent the ORFs sharing a similar codon usage pattern. The bar denotes a difference of 0.1 unit.

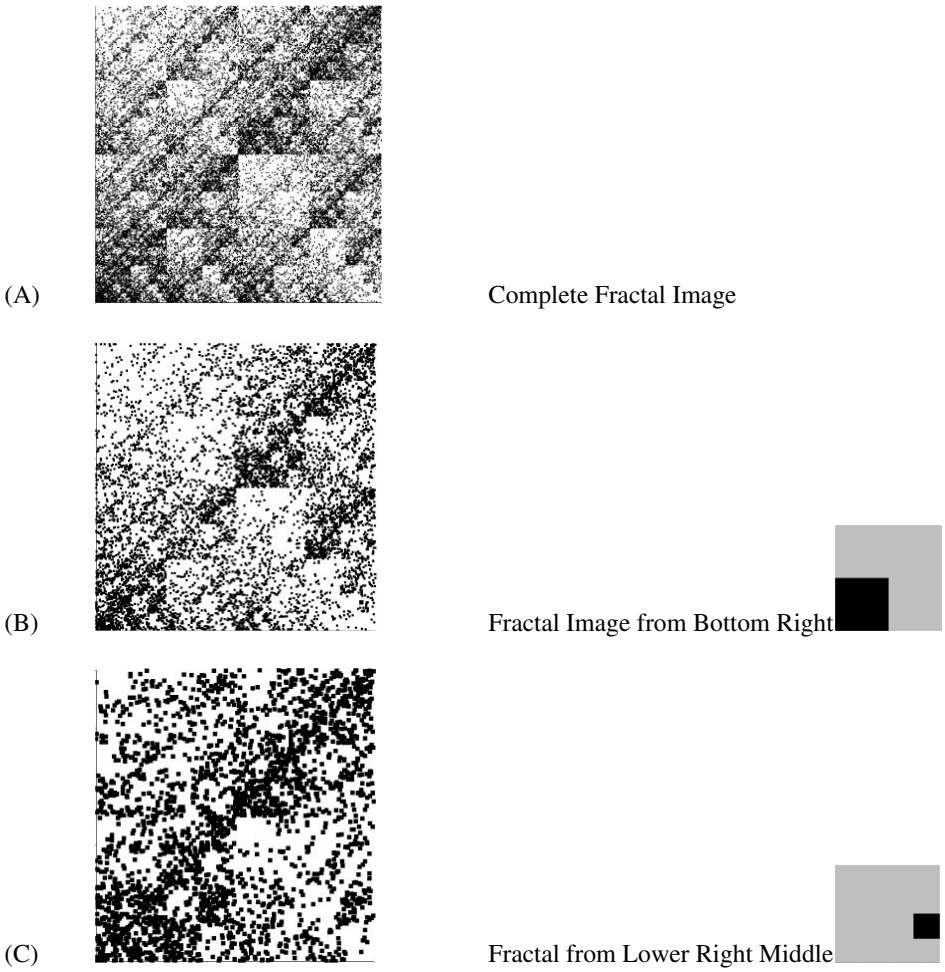

Fig. 4

Fractal pattern of complete genome (A) and areas focused in more details (B and C)

170 and ORF-65 showed any degree of similarity to ORFs from other systems. It is also interesting to note that five of the nine putative ORFs that did not primarily occupy the bottom left quadrant of their gene landscape were amongst the smallest ORFs (77, 65, 49, 77 and 52 amino acids in length, respectively) that have been predicted for this genome. Although genome landscaping has been shown as a means of studying potential genome duplication events, the role of gene landscaping in bioinformatical analysis is unclear. Nevertheless, it is interesting to note that the majority of the landscapes for ORFs from this particular bacteriophage occupied the same quadrant on the landscape. The role, if any, of applying gene landscaping studies will become clearer when genes from additional genomes are analysed.

MADCA-BORU analysis

Methods similar to MADCA-BORU, e.g. Euclidean distance calculations depend on the proteins encoded by different genes having relatively similar amino acid composition, and can be applied to analysis of genes with similar functions. In the case of ORFs from a complete genome, where this is not the case, there must be some degree of compensation for amino acid composition – hence the use of the MADCA-BORU values for this work.

The tree of bacteriophage Φadh genes (ORFs) of known function as assessed by the use of MADCA-BORU is shown in Fig. 3. However, the trees of all ORFs present or those of a combination of genes and ORFs with similarity to those in the databases showed similar distribution patterns. Due to the

number of sequences that clustered close together the branch names have been removed and replaced with group names. Branch names for each group on the tree are listed in the legend to Fig. 3. The tree contains four distinct groups. Groups 1 and 2 contain a mixture of ORFs with a known function and hypothetical ORFs. Group 3 represents only two ORFs, but both are known to be functional. Group 4 contains five ORFs, and none of them has been shown to have any function within the phage and any significant similarity to either genes of a known function, or even to other hypothetical ORFs. Furthermore, all of these ORFs are short, and none is long enough to allow analysis by either MRI or Nc. In addition, only one of the five ORFs in this group (ORF-55) had a gene landscape lying in the bottom left quadrant.

Fractal analysis

Fractal analysis demonstrates that there is no obvious DNA positional bias, as opposed to over all compositional bias within the genome, with the complete fractal picture showing a strong similarity to areas within it (Fig. 4). Thus the fractal picture shows what appears to be a Julia style distribution (Julia, 1918) – with Julia sets being defined as being self-similar (i.e. zooming in on any particular area of the fractal pattern produces a pattern similar to the whole fractal pattern).

The ORFs in the bacteriophage Φadh genome make use of a restricted number of codons and in areas with the potential to encode ORFs there is a bias in favour of pyrimidines on the coding strand. Based on their codon usage pattern, as opposed to relative number of codons being used, the ORFs could be split into one of four clusters by MADCA-BORU analysis. Furthermore, the fractal analysis of the genome showed that the nucleotides were distributed in a Julia-style fractal pattern.

Note of the Editor-in-Chief. The bacteriophage Φadh is not listed among viruses (virus species) in the presently valid taxonomy of viruses (van Regenmortel MHV, Fauquet CM, Bishop DHL: *Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses.* Academic Press, San Diego-San-Francisco-New York-Boston-London-Sydney-Tokyo, 2000).

References

Altermann E, Klein JR, Henrich B (1999): Primary structure and features of the genome of the *Lactobacillus gasseri* temperate bacteriophage Φadh. *Gene* **236**, 333–346.

Engel G, Altermann E, Klein JR, Henrich B (1998) Structure of a genome region of the *Lactobacillus gasseri* temperate bacteriophage Φadh covering a repressor gene and cognate promoter. *Gene* **210**, 61–70.

Felsenstein J (1989): PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166.

Freire-Picos MA, Gonzalez-Siso MI, Rodriguez-Belmonte E, Rodriguez-Torres AM, Ramil E, Cerdan ME (1994): Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes. *Gene* **139**, 43–49.

Fremaux C, de Antoni GL, Raya RR, Klaenhammer TR (1993): Genetic organization and sequence of the region encoding integrative functions from *Lactobacillus gasseri* temperate bacteriophage Φadh. *Gene* **126**, 61–66.

Gatherer D, McEwan NR (1997): Small regions of preferential codon usage and their effect on overall codon bias – the case of the *plp* gene. *Biochem. Mol. Biol. Int.* **43**, 107–114.

Gouy M, Gautier C (1982): Codon usage in bacteria – correlation with gene expressivity. *Nucleic Acids Res.* **10**, 7055–7074.

Henrich B, Binishofer B, Bläsi U (1999): Primary Structure and Functional Analysis of the Lysis Genes of *Lactobacillus gasseri* Bacteriophage Φadh. *J. Bacteriol.* **177**, 723–732.

Itoh T, Fujimoto Y, Kawai Y, Toba T, Saito T (1995): Inhibition of food-borne pathogenic bacteria by bacteriocins from *Lactobacillus gasseri*. *Lett. Appl. Microbiol.* **21**, 137–141.

Julia GM (1918): Mémoire sur l'itération des fonctions rationnelles. *J. Math. Pure et Appl*. **8**, 47–245.

Lobry JR (1999): Genomic Landscapes. *Microbiol. Today* **26**, 164–165.

McEwan NR, Gatherer D (1998): The Mutational Response Index and Codon Bias in Genes from a *Frankia nif* Operon. *Theor. Appl. Genet.* **96**, 716–718.

McEwan NR, Gatherer D (1999): Codon indices as a predictor of gene functionality in a *Frankia* operon. *Can. J. Bot.* **77**, 1287–1292.

Metchnikoff E (1907): In *The Prolongation of Life – Optimistic Studies*. W. Heinemann, London, pp. 161–183.

Page RDM (1996): TREEVIEW, An application to display phylogenetic trees on personal computers. *Comp. Appl. Biosci.* **12**, 357–358.

Raya RR, Kleeman EG, Luchansky JB, Klaenhammer TR (1989): Characterization of a temperate bacteriophage Φadh and plasmid transduction in *Lactobacillus acidophilus* ADH. *Appl. Environ. Microbiol.* **55**, 2206–2213.

Raya RR, Fremaux C, de Antoni GL, Klaenhammer TR (1992): Site-Specific Integration of the Temperate Bacteriophage Φadh into the *Lactobacillus gasseri* Chromosome and Molecular Characterization of the Phage (*attP*) and Bacterial (*attB*) Attachment Sites. *J. Bacteriol.* **174**, 5584–5592.

Sharp PM, Li W-H (1987): The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295

Tahara T, Yoshioka S, Utsumi R, Kanatani K (1997): Isolation and partial characterization of bacteriocins produced by *Lactobacillus gasseri* JCM 2124. *FEMS Microbiol. Lett.* **148**, 97–100.

Wright F (1990): The 'effective number of codons' used in a gene. *Gene* **87**, 23–29.