

Table S2. Classification and regression tree analysis results using 4 oligonucleotide frequency lengths over the entire virtual coding strand to classify organisms into two temperature optimum ranges, hyperthermophile versus non-hyperthermophile (hyper vs. non-hyper) and mesophile versus non-mesophile (meso vs. non-meso). Tetramers provide the most effective classification scheme considering both models built. Model-with-all is the number of misclassifications when building a model using all 195 genomes. Leave-one-out is the number of mispredictions when classifying an organism against a model built using the other 194 genomes.

Motif length (L-mer)	Model-with-all		Leave-one-out	
	Hyper vs. non-hyper	Meso vs. non-meso	Hyper vs. non-hyper	Meso vs. non-meso
2-mer	5 (2.6%)	8 (4.1%)	10 (5.1%)	24 (12.3%)
3-mer	4 (2.1%)	6 (3.1%)	12 (6.2%)	14 (7.2%)
4-mer	2 (1.0%)	9 (4.6%)	8 (4.1%)	14 (7.2%)
5-mer	2 (1.0%)	9 (4.6%)	13 (6.7%)	23 (11.8%)