Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Differences in dinucleotide frequencies of thermophilic genes encoding water soluble and membrane proteins

Hiroshi NAKASHIMA†, Yuka KURODA

(*Department of Clinical Laboratory Science, Graduate Course of Medical Science and Technology, School of Health Sciences, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa 920-0942, Japan*)

†E-mail: naka@kenroku.kanazawa-u.ac.jp

Received Sept. 13, 2010; Revision accepted May 5, 2011; Crosschecked May 9, 2011

Abstract: The occurrence frequencies of the dinucleotides of genes of three thermophilic and three mesophilic species from both archaea and eubacteria were investigated in this study. The genes encoding water soluble proteins were rich in the dinucleotides of purine dimers, whereas the genes encoding membrane proteins were rich in pyrimidine dimers. The dinucleotides of purine dimers are the counterparts of pyrimidine dimers in a double-stranded DNA. The purine/pyrimidine dimers were favored in the thermophiles but not in the mesophiles, based on comparisons of observed and expected frequencies. This finding is in agreement with our previous study which showed that purine/pyrimidine dimers are positive factors that increase the thermal stability of DNA. The dinucleotides AA, AG, and GA are components of the codons of charged residues of Glu, Asp, Lys, and Arg, and the dinucleotides TT, CT, and TC are components of the codons of hydrophobic residues of Leu, Ile, and Phe. This is consistent with the suitabilities of the different amino acid residues for water soluble and membrane proteins. Our analysis provides a picture of how thermophilic species produce water soluble and membrane proteins with distinctive characters: the genes encoding water soluble proteins use DNA sequences rich in purine dimers, and the genes encoding membrane proteins use DNA sequences rich in pyrimidine dimers on the opposite strand.

Key words: Water soluble and membrane proteins, Purine/pyrimidine dimers, Thermophilic and mesophilic species, Dinucleotide frequencies

doi:10.1631/jzus.B1000331 **Document code:** A **CLC number:** Q78

1 Introduction

The G+C content of bacterial genomes varies among species from 25% to 75%, but is relatively constant within a bacterial genome (Muto and Osawa, 1987; Lawrence and Ochman, 1997). The nucleotide sequences of genes of bacterial genomes have species-specific dinucleotide compositions (Karlin and Burge, 1995; Karlin *et al.*, 1997; Nakashima *et al.*, 1998). Comparative studies of the DNA and protein sequences of thermophilic and mesophilic species have revealed differences in their compositions. The synonymous codon usage in genes of thermophiles is different from that of mesophiles (Lynn *et al.*, 2002). Kawashima *et al.* (2000) reported that in archaea a simple combination of purine (R) and pyrimidine (Y) dinucleotides, RR+YY-RY-YR, is linearly correlated with optimal growth temperature (OGT). An increased frequency of purine nucleotides in the coding strands contributes to thermostability (Paz *et al.*, 2004). It has been reported that a simple summation of the purines adenine and guanine (A+G) is correlated with OGT (Lambros *et al.*, 2003; Zeldovich *et al.*, 2007). Studies of thermophilic and mesophilic proteins have shown differences in their amino acid compositions (Kumar *et al.*, 2000; Kreil and Ouzounis, 2001; Farias and Bonato, 2003; Yokota *et al.*, 2006; Zhou *et al.*, 2008).

We previously reported that the ten symmetrical

components of the dinucleotide compositions of genes encoding water soluble proteins showed a linear relationship with OGT based on regression analysis (Nakashima *et al.*, 2003). The purine/pyrimidine dimers were positive, but purine-pyrimidine or pyrimidine-purine dimers were roughly negative factors for the thermal stability of DNA. The dinucleotide AA pairs with TT and we cannot distinguish AA from TT in a double-stranded DNA. The dinucleotide AT pairs with AT. Therefore, ten symmetrical components are enough to study the character of a double-stranded DNA. When we consider a coding sequence, it is important to recognize on which strand the gene is located. In this case, we have to consider 16 kinds of dinucleotides.

It has been estimated that more than a quarter of all known proteins are membrane proteins (Anson, 2009). These proteins have different amino acid compositions from water soluble proteins. Apolar amino acid residues are suitable for membrane proteins because such proteins have membrane-spanning regions which have hydrophobic characteristics. As there is a difference in amino acid composition between water soluble and membrane proteins, the dinucleotide compositions of their genes must be different. This raises the issue of how the species prepare two different kinds of DNA sequences. To address this question, we investigated the dinucleotide compositions of membrane proteins from both thermophilic and mesophilic species and compared them with those of water soluble proteins.

2 Materials and methods

2.1 Sequence retrieval

Three thermophilic archaea (*Sulfolobus tokodaii* (Kawarabayasi *et al*., 2001), *Archaeoglobus fulgidus* (Klenk *et al.*, 1997), *Methanopyrus kandleri* (Slesarev *et al.*, 2002)), three thermophilic eubacteria (*Thermoanaerobacter tengcongensis* (Bao *et al.*, 2002), *Thermotoga maritima* (Nelson *et al.*, 1999), *Thermus thermophilus* HB8 (GenBank: AP008226.1)), three mesophilic archaea (*Methanosphaera stadtmanae* (Fricke *et al*., 2006), *Methanocorpusculum labreanum* (Anderson *et al*., 2009), *Halobacterium* sp. NRC-1 (Ng *et al*., 2000)), and three mesophilic eubacteria (*Haemophilus influenzae* Rd KW20 (Fleischmann *et al.*, 1995), *Escherichia coli* K12 MG1655 (Blattner *et al.*, 1997), *Pseudomonas aeruginosa* PA01 (Stover *et al.*, 2000)) were surveyed in this study. The species were selected arbitrarily, taking into consideration only the coverages of a wide range of genomic G+C contents. Their genome sequences were retrieved from the web ftp site (ftp://ftp.ncbi.gov/genomes/) of the National Center for Biotechnology Information (NCBI). The protein-coding nucleotide sequences and amino acid sequences were retrieved from NCBI as ffn and faa files.

2.2 Selection of genes encoding water soluble and membrane proteins

The proteins were classified as water soluble or membrane proteins according to the annotation of the Genome to Protein Structure and Function (GTOP) database (Kawabata *et al*., 2002). The SOSUI program (Hirokawa *et al.*, 1998) was used in the GTOP database to predict the transmembrane regions. Those proteins without transmembrane regions were considered to be water soluble proteins. Proteins which had more than two transmembrane regions were employed in the calculation of the dinucleotide compositions of genes for membrane proteins. This is because if a protein has a signal peptide it might be counted as a transmembrane region. The membrane proteins were divided into 100 sections and one protein was randomly selected from each section. The water soluble proteins were similarly selected. The water soluble and membrane proteins were examined for their amino acid sequence similarities using the BLAST program (Altschul *et al*., 1990). Those proteins which had more than 30% sequence identity between water soluble proteins or between membrane proteins were replaced, to keep the sequence identity below 30%. Proteins longer than 100 residues, and their corresponding genes, were employed.

2.3 Calculation of expected dinucleotide composition

The expected dinucleotide composition was calculated as the product of the mononucleotide composition for each gene. The averages of the expected dinucleotide compositions for 100 genes encoding water soluble and membrane proteins were calculated. Then the ratios of the average of the observed to the expected dinucleotide compositions were calculated.

3 Results

3.1 Dinucleotide composition

The average dinucleotide compositions of the genes for 100 water soluble proteins and 100 membrane proteins in 12 species are listed in Table 1 with their genomic G+C contents. In *T. maritima*, dinucleotides such as AA, GA, and AG were enriched in the water soluble protein-coding genes, whereas TT, TC, and CT were enriched in the membrane protein-coding genes. It is interesting that AA, GA, and AG are purine dimers and TT, TC, and CT are pyrimidine dimers, and they are counterparts in a double helix DNA. To show the difference in dinucleotide composition more clearly, the ratios of the dinucleotide compositions of water soluble protein-coding genes to those of membrane protein-coding genes were calculated. In *T. maritima*, the three highest ratios of dinucleotide compositions were AA (1.56=11.20/7.20), AG (1.50=8.55/5.69), and GA (1.46=10.99/7.54). This result indicated that the genes encoding water soluble proteins were rich in AA, AG, and GA dinucleotides compared to the genes encoding membrane proteins. The three lowest ratios of dinucleotide compositions were TT (0.58=5.92/10.23), CT (0.68=5.38/7.92), and TC (0.72=6.71/9.29) in *T. maritima*. This result indicated that the dinucleotides TT, CT, and TC were frequently observed in the genes encoding membrane proteins compared to the genes encoding water soluble proteins (Table 2). The genes for water soluble proteins were biased toward purine dimers such as GA, AA, and AG, and the genes for membrane proteins were biased toward pyrimidine dimers such as TC, TT, and CT. This trend was observed in both the thermophiles and the mesophiles.

The occurrence frequency of dinucleotides was dependent on G+C content. For example, *H. influenzae* with a genomic G+C content of 38.1% showed a higher occurrence frequency for the dinucleotides AA and TT, and a lower occurrence frequency for the dinucleotides CC and GG in the genes for both water soluble and membrane proteins. *P. aeruginosa*, with a genomic G+C content of 66.6%, showed a lower occurrence frequency for the dinucleotides AA and TT, and a higher occurrence frequency for the dinucleotides CC and GG.

To analyze the difference between thermophilic and mesophilic genes, the sums of purine/pyrimidine dimers were calculated. The sum of purine dimers for the water soluble protein-coding genes of *T. maritima* was 37.39% and that for the membrane protein-coding genes was 26.70% (Table 1). So the deviation of the two sets of genes was 10.69%. The sum of pyrimidine dimers for the water soluble protein-coding genes of *T. maritima* was 21.97% and that for the membrane protein-coding genes was 31.58%. In this case, the deviation was 9.61%. The corresponding deviations were 4.99% and 4.38% in *E. coli*. Thus, the deviation of the sum of purine/pyrimidine dimers between the water soluble and the membrane proteins-coding genes was generally larger in the thermophiles than in the mesophiles. The larger deviation implied that the protein-coding genes in thermophiles are more biased towards purine/pyrimidine dimers than those in mesophiles.

3.2 Favorable purine/pyrimidine dimers in thermophiles

The ratios of the observed to the expected dinucleotide compositions among species were calculated. Ratios greater than 1.1 were considered favorable and those less than 0.9 were considered unfavorable. To simplify the result, only purine/pyrimidine dimers were considered here. In *T. maritima*, TC, GA, CT, TT and AA were favorable dinucleotides for the genes encoding water soluble proteins and no unfavorable dinucleotides were observed. The dinucleotides TC, GA, AA, CT, TT, and GG were favorable and the dinucleotide CC was unfavorable for the genes encoding membrane proteins in *T. maritima*. Thus, there were five favorable purine/pyrimidine dimers in the genes encoding water soluble proteins, and six in those encoding membrane proteins. One pyrimidine dimer was observed as unfavorable for the genes encoding membrane proteins. The favorable dinucleotides were almost identical for the genes encoding both protein types. This is consistent with the result of Karlin's group that the dinucleotide relative abundance values of different DNA sequences from the same organism are generally much more similar to each other than those from different organisms (Karlin and Burge, 1995; Karlin *et al*., 1997). The dinucleotide relative abundance values from Karlin's group correspond to the ratios of the observed to the expected dinucleotide compositions in our study.

Table 1 Average dinucleotide compositions of 100 genes encoding water soluble and membrane proteins

Species	G+C (%)	Dinucleotide composition (%)																Sum (%)	
		AA	TT	AG	CT	GA	TC	GG	CC	AC	GT	CA	TG	AT	TA	GC	CG	RR	YY
Thermophilic archaea																			
S. tokodaii	32.8																		
Soluble		13.00	9.29	8.69	5.03	8.26	4.02	5.04	2.64	4.40	5.02	4.43	5.65	9.79	10.13	2.84	1.77	34.99	20.98
Membrane		8.92	14.16	5.87	6.94	4.52	5.84	3.99	2.84	4.29	5.18	4.73	5.38	11.03	11.88	3.00	1.43	23.30	29.78
A. fulgidus	48.6																		
Soluble		8.58	6.51	9.42	5.57	10.30	4.93	8.34	4.12	4.89	4.59	5.74	7.00	5.39	3.60	6.27	4.75	36.64	21.13
Membrane		5.91	9.67	6.02	8.02	6.44	7.53	7.03	5.17	4.55	4.89	5.76	7.07	6.06	4.35	6.63	4.90	25.40	30.39
M. kandleri	61.2																		
Soluble		4.52	2.77	7.42	4.66	10.13	6.59	10.05	7.06	6.09	6.58	4.09	5.40	3.76	3.01	6.99	10.88	32.12	21.08
Membrane		3.03	3.93	4.79	6.72	6.75	7.98	10.05	7.64	5.73	7.19	4.09	6.31	4.18	3.83	7.45	10.33	24.62	26.27
Thermophilic eubacteria																			
T. tengcongensis	37.6																		
Soluble		13.40	8.94	9.19	4.83	9.02	3.24	6.16	2.86	4.19	4.48	5.08	6.80	8.29	7.54	4.23	1.75	37.77	19.87
Membrane		9.67	13.02	6.65	6.47	5.98	4.36	5.69	3.04	4.30	4.68	5.21	7.07	8.79	8.49	4.81	1.77	27.99	26.89
T. maritima	46.2																		
Soluble		11.20	5.92	8.55	5.38	10.99	6.71	6.65	3.96	5.85	5.19	5.73	6.22	5.75	3.38	3.55	4.97	37.39	21.97
Membrane		7.20	10.23	5.69	7.92	7.54	9.29	6.27	4.14	4.78	5.98	5.44	7.14	6.17	3.62	3.96	4.65	26.70	31.58
T. thermophilus	69.5																		
Soluble		2.94	2.67	6.17	6.83	6.75	6.09	15.32	14.60	4.84	3.49	4.71	4.74	1.77	1.25	9.25	8.58	31.18	30.19
Membrane		1.77	3.95	3.79	10.10	4.07	8.88	14.08	15.53	4.00	3.73	4.12	5.40	1.80	1.35	9.40	8.03	23.71	38.46
Mesophilic archaea																			
M. stadtmanae	27.6																		
Soluble		15.76	8.35	6.75	3.93	7.31	3.42	3.43	1.92	5.24	4.74	6.53	7.11	12.26	10.38	2.34	0.53	33.25	17.62
Membrane		11.73	12.33	5.48	4.45	4.91	3.99	3.50	1.62	4.79	4.99	6.22	6.31	13.96	13.06	2.28	0.38	25.62	22.39
M. labreanum	50.0																		
Soluble		8.43	4.83	5.76	5.08	8.75	7.48	6.69	6.01	5.78	4.86	6.45	6.18	6.88	3.14	6.00	7.68	29.63	23.40
Membrane		5.47	8.06	3.82	6.75	6.64	9.20	7.02	5.50	4.28	5.74	5.31	7.19	7.73	3.81	6.04	7.44	22.95	29.51
Halobacterium	67.9																		
Soluble		2.51	1.72	4.49	4.14	8.66	6.50	9.14	9.00	8.18	5.29	5.32	4.33	2.48	1.08	10.99	16.17	24.80	21.36
Membrane		1.82	3.00	2.79	5.66	5.18	8.17	10.06	7.90	5.88	7.54	4.62	6.25	2.76	1.52	11.56	15.29	19.85	24.73
Mesophilic eubacteria																			
H. influenzae	38.1																		
Soluble		13.15	10.39	5.39	4.27	6.14	4.53	4.25	3.01	4.87	5.35	6.27	7.25	8.88	6.68	5.36	4.20	28.93	22.20
Membrane		9.01	14.92	4.01	5.21	4.30	4.84	4.85	2.86	4.22	5.78	5.67	7.91	9.57	7.76	5.47	3.63	22.17	27.83
E. coli	50.8																		
Soluble		7.72	5.89	4.94	5.28	6.74	5.25	6.55	5.23	5.69	5.55	6.25	8.00	6.45	4.00	8.53	7.92	25.95	21.65
Membrane		5.25	8.85	3.44	6.50	4.79	5.86	7.48	4.82	4.46	6.63	5.23	9.42	6.82	4.64	8.62	7.19	20.96	26.03
P. aeruginosa	66.6																		
Soluble		3.36	2.30	5.13	5.98	6.51	5.68	8.66	10.06	5.65	4.06	6.03	6.20	3.24	1.40	13.24	12.49	23.66	24.02
Membrane		2.09	2.97	3.38	8.39	4.62	7.14	9.18	9.67	4.37	4.87	5.18	8.24	3.49	1.36	13.59	11.45	19.27	28.17

Table 2 List of dinucleotides frequently observed in the genes encoding water soluble and membrane proteins

Species	Dinucleotide	
	Water soluble proteins	Membrane proteins
Thermophilic archaea		
S. tokodaii	GA, AG, AA	TT, TC, CT
A. fulgidus	GA, AG, AA	TC, TT, CT
M. kandleri	AG, GA, AA	CT, TT, TA
Thermophilic eubacteria		
T. tengcongensis	GA, AA, AG	TT, TC, CT
T. maritima	AA, AG, GA	TT, CT, TC
T. thermophilus	GA, AA, AG	CT, TT, TC
Mesophilic archaea		
M. stadtmanae	GA, CG, AA	TT, TA, TC
M. labreanum	AA, AG, AC	TT, CT, TC
Halobacterium	GA, AG, AC	TT, TG, GT
Mesophilic eubacteria		
H. influenzae	AA, GA, AG	TT, CT, TA
E. coli	AA, AG, GA	TT, CT, GT
P. aeruginosa	AA, AG, GA	CT, TG, TT

In *E. coli*, the dinucleotides AA and TT were favorable for the genes encoding both water soluble and membrane proteins. The dinucleotides AG, CC, and GG were unfavorable for the genes encoding water soluble proteins and the dinucleotides AG, CC, TC, and GA were unfavorable for the genes encoding membrane proteins. So, for both protein types, there were two favorable purine/pyrimidine dimers. There were three unfavorable purine/pyrimidine dimers for the genes encoding water soluble proteins and four for those encoding membrane proteins (Table 3). There are eight purine/pyrimidine dimers in total. In *T. thermophilus*, almost all purine/pyrimidine dimers were classed as favorable. This indicates that purine/pyrimidine dimers were favorable in the thermophiles, except *M. kandleri*, but not in the mesophiles.

3.3 Sum of purine/pyrimidine dimers along a nucleotide sequence

The dinucleotides TT, TC, and CT were frequently observed in the genes encoding membrane proteins. To analyze the occurrence of pyrimidine dimers, the sum of the dinucleotides CT+TC+TT+CC was counted in a frame of 30 nucleotides, sliding the frame without overlapping along a nucleotide sequence. Similarly, the sum of the dinucleotides AG+GA+AA+GG was counted. The plot of the sum of purine/pyrimidine dimers along the nucleotide sequence which encodes the ABC transporter permease protein of *T. maritima* is shown in Fig. 1. The peaks of the sum of pyrimidine dimers correspond to the transmembrane regions and the local minima of the sum of purine dimers. Thus the sum of pyrimidine dimers corresponds to the transmembrane region. This result suggests that multi-spanning transmembrane proteins have more pyrimidine dimers than single-spanning transmembrane proteins. The dinucleotides CT, TC, and TT were frequently observed as components of codons such as Leu (CTN, TTA, TTG), Ser (TCN), and Phe (TTC, TTT) in transmembrane regions.

Table 3 Numbers of favorable and unfavorable purine/pyrimidine dimers in the genes encoding water soluble and membrane proteins

Species	Water soluble proteins		Membrane proteins	
	n_f	n_u	n_f	n_u
Thermophilic archaea				
S. tokodaii	6	0	4	0
A. fulgidus	5	0	7	0
M. kandleri	2	2	3	2
Thermophilic eubacteria				
T. tengcongensis	7	1	6	1
T. maritima	5	0	6	1
T. thermophilus	8	0	7	0
Mesophilic archaea				
M. stadtmanae	3	0	1	3
M. labreanum	3	2	4	2
Halobacterium	2	5	2	4
Mesophilic eubacteria				
H. influenzae	2	4	3	4
E. coli	2	3	2	4
P. aeruginosa	4	2	4	4

n_f: number of favorable purine/pyrimidine dimers; n_u: number of unfavorable purine/pyrimidine dimers

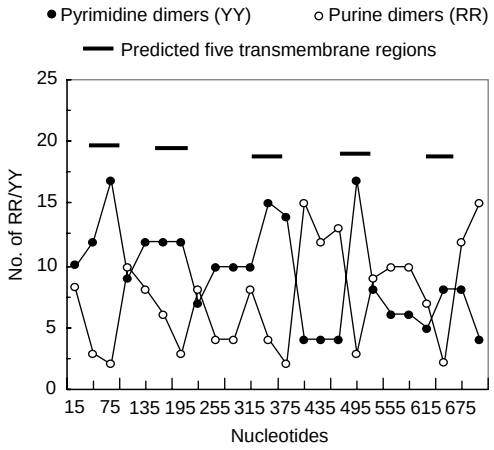

Fig. 1 A plot of the sum of the purine/pyrimidine dimers in a frame of 30 nucleotides, sliding the frame without overlapping along a nucleotide sequence
The nucleotide sequence is the ABC transporter permease protein of *T. maritima*

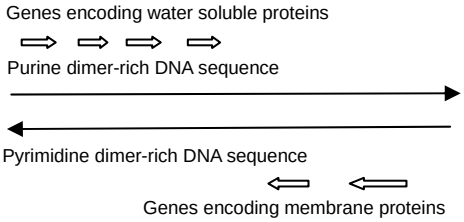

Fig. 2 Schematic picture showing the genes encoding water soluble and membrane proteins
Water soluble and membrane proteins are likely to sit in a series on a DNA strand

4 Discussion

As expected, the dinucleotide compositions of genes encoding water soluble and membrane proteins were different (Tables 1 and 2). The genes encoding water soluble proteins were rich in the dinucleotides AG, AA, and GA, and their counterparts CT, TT, and TC were abundant in the genes encoding membrane proteins. The above purine/pyrimidine dimers were favorable in the genes from the thermophiles, but not in the genes from the mesophiles (Table 3). We tried to understand how the organisms prepare two kinds of different DNA sequences. The organisms use a simple strategy: the genes for water soluble and membrane proteins use DNA sequences on different DNA strands (Fig. 2). The protein-coding genes from the thermophiles were richer in purine/pyrimidine dimers than those from the mesophiles. This is consistent with our previous study which suggested that the purine/pyrimidine dimers were positive factors that increased the thermal stability of DNA (Nakashima *et al.*, 2003). The dinucleotides AA, AG, and GA are components of the codons of charged residues of Glu, Asp, Lys, and Arg, and these residues are known to stabilize proteins at higher temperatures (Nakashima *et al*., 2003). The dinucleotides TT, CT, and TC are components of the codons of hydrophobic residues of Leu, Ile, and Phe. This is consistent with the suitability of these amino acid residues for membrane proteins.

We showed the results from three thermophilic and three mesophilic species of both archaea and eubacteria in this study. We examined the dinucleotide compositions of genes from other thermophilic and mesophilic species, and obtained a similar trend for the dinucleotide composition. The dinucleotide composition is thought to consist of two parts: one is attributable to mononucleotide composition, and the other is a deviation from expectation given by the multiplication of mononucleotide contents. In a double-stranded DNA, the amount of adenine is equal to the amount of thymine and the amount of guanine is equal to the amount of cytosine. This is known as Chargaff's first parity rule (Chargaff *et al*., 1951; 1952). This rule also applies to single-stranded DNA and is called Chargaff's second parity rule (Karkas *et al.*, 1968; Runder *et al*., 1968). Mitchell and Bridge (2006) tested Chargaff's second parity rule over 3400 genomic sequences and the validity of this rule has been confirmed for genome sequences from archaea, eubacteria, eukaryotes, and viruses. Therefore, the mononucleotide composition is represented simply as G+C content. This is why we selected species showing a wide range of G+C contents. The present study indicated that the character of the dinucleotide composition held for genes of a wide range of G+C contents.

The occurrence frequencies of the dinucleotides GG and CC in the genes were low compared to those of other purine/pyrimidine dimers. In the thermophiles, the occurrence frequency of the dinucleotide GG was higher in the genes encoding water soluble proteins than in those encoding membrane proteins. However, the opposite trend was found in the mesophiles (Table 1). The dinucleotide GG is a component of the codon of the Gly residue (GGN). The character of Gly might be related to the above occurrence

frequency of GG. The degree of localization of the dinucleotide CC in the transmembrane regions was low compared to that of other pyrimidine dimers. The dinucleotide CC is a component of the codon of the Pro residue (CCN). Transmembrane regions are composed of α-helices, and the Pro residue is a strong helix-breaker (Chou and Fasman, 1978), so this might be the reason why the dinucleotide CC is not favored in transmembrane regions.

The DNA strand in which the genes for water soluble proteins were located was different from the strand carrying the genes for membrane proteins. Motivated by this result, we investigated the distributions of genes encoding water soluble and membrane proteins. The genes were divided into two types, so four types of gene pairs were possible. The occurrence of the four different types of gene pairs was counted whenever the two genes were located successively on an identical strand. Then the observed number of gene pairs was divided by the calculated number to obtain a ratio, i.e., observed/calculated. The calculated number of gene pairs was estimated by the product of the frequency of the genes. This procedure was followed for both strands separately. The two strands were represented by a top strand and a bottom strand. The ratios (observed/calculated) of the four types of gene pairs on both strands in 12 species are listed in Table 4. The numbers of genes encoding water soluble and membrane proteins on both strands are also listed in Table 4. The ratios of the gene pairs

Table 4 Occurrence ratios (observed/calculated) of pairs between water soluble and membrane proteins on an identical strand

Species	Strand	No. of proteins		Ratio of protein pairs			
		Soluble	Membrane	S-S	S-M	M-S	M-M
Thermophilic archaea							
S. tokodaii	Top	1035	383	1.11	0.79	0.79	1.34
	Bottom	1065	343	1.12	0.64	0.69	1.88
A. fulgidus	Top	913	265	1.03	0.82	0.79	1.95
	Bottom	923	306	1.06	0.78	0.75	1.85
M. kandleri	Top	671	172	1.08	0.64	0.75	2.16
	Bottom	676	168	1.03	0.78	0.82	2.08
Thermophilic eubacteria							
T. tengcongensis	Top	919	333	1.03	0.93	0.90	1.24
	Bottom	954	382	1.10	0.80	0.81	1.34
T. maritima	Top	723	280	1.08	0.80	0.77	1.59
	Bottom	634	209	1.03	0.85	0.88	1.52
T. thermophilus	Top	674	217	1.00	0.92	0.86	1.62
	Bottom	834	248	1.04	0.78	0.92	1.54
Mesophilic archaea							
M. stadtmanae	Top	586	163	1.06	0.78	0.80	1.79
	Bottom	616	169	1.06	0.85	0.69	1.83
M. labreanum	Top	667	199	1.12	0.73	0.67	1.71
	Bottom	650	223	1.11	0.70	0.73	1.76
Halobacterium	Top	973	270	1.07	0.82	0.79	1.55
	Bottom	1059	303	1.01	0.79	0.95	1.76
Mesophilic eubacteria							
H. influenzae	Top	656	194	1.01	0.79	0.92	1.83
	Bottom	659	200	1.01	0.86	0.89	1.70
E. coli	Top	1448	593	1.04	0.87	0.91	1.30
	Bottom	1596	546	1.04	0.93	0.84	1.37
P. aeruginosa	Top	2010	732	1.05	0.78	0.90	1.53
	Bottom	2065	759	1.04	0.90	0.84	1.42

S-S: soluble-soluble; S-M: soluble-membrane; M-S: membrane-soluble; M-M: membrane-membrane

corresponding to two membrane proteins were greater than 1, indicating that they were favorable in all species, whereas those of gene pairs corresponding to a water soluble and a membrane protein (and the reverse) were less than 1, indicating that these were unfavorable in all species. This result indicates that genes encoding membrane proteins are likely to sit in a series on a DNA strand. In *E. coli*, the genes encoding membrane proteins such as the cytochrome o ubiquinol oxidase subunit, the nicotinamide adenine dinucleotide (NADH) ubiquinone oxidoreductase subunit, and the adenosine triphosphate (ATP) synthase subunit, were located successively as components of operons. This result agrees with the empirical knowledge that functionally-related genes are clustered in an operon. This is also consistent with Chargaff's cluster rule that purine/pyrimidine nucleotides tend to cluster in a DNA sequence (Forsdyke and Mortimer, 2000).

5 Conclusions

The genes encoding water soluble proteins use DNA sequences rich in purine dimers, and the genes encoding membrane proteins use sequences rich in pyrimidine dimers on the opposite strand. The dinucleotides AA, AG, and GA are components of the codons of charged residues of Glu, Asp, Lys, and Arg, and the dinucleotides TT, CT, and TC are components of the codons of hydrophobic residues of Leu, Ile, and Phe. This is consistent with the suitabilities of the different amino acid residues for water soluble and membrane proteins. The protein-coding genes from the thermophiles were richer in purine/pyrimidine dimers than those from the mesophiles.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.*, **215**(3):403-410. [doi:10.1016/S0022-2836(05)80360-2]

Anderson, I., Ulrich, L.E., Lupa, B., Susanti, D., Porat, I., Hooper, S.D., Lykidis, A., Sieprawska-Lupa, M., Dharmarajan, L., Goltsman, E., *et al*., 2009. Genomic characterization of methanomicrobiales reveals three classes of methanogens. *PLoS One*, **4**(6):1-9. [doi:10.1371/journal.pone.0005797]

Anson, L., 2009. Membrane protein biophysics. *Nature*, **459**(7245):343. [doi:10.1038/459343a]

Bao, Q., Tian, Y., Li, W., Xu, Z., Xuan, Z., Hu, S., Dong, W., Yang, J., Chen, Y., Xue, Y., *et al*., 2002. A complete sequence of the *T. tengcongensis* genome. *Genome Res*., **12**(5):689-700. [doi:10.1101/gr.219302]

Blattner, F.R., Plunkett, G.III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., *et al*., 1997. The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**(5331):1453-1462. [doi:10.1126/science.277.5331.1453]

Chargaff, E., Lipshitz, R., Green, C., Hodes, M.E., 1951. The composition of the deoxyribonucleic acid of salmon sperm. *J. Biol. Chem*., **192**(1):223-230.

Chargaff, E., Lipshitz, R., Green, C., 1952. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J. Biol. Chem*., **195**(1):155-160.

Chou, P.Y., Fasman, G.D., 1978. Empirical predictions of protein conformation. *Ann. Rev. Biochem*., **47**(1):251-276. [doi:10.1146/annurev.bi.47.070178.001343]

Farias, S.T., Bonato, M.C.M., 2003. Preferred amino acids and thermostability. *Genet. Mol. Res.*, **2**(4):383-393.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.*, 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**(5223):496-512. [doi:10.1126/science.7542800]

Forsdyke, D.R., Mortimer, J.R., 2000. Chargaff's legacy. *Gene*, **261**(1):127-137. [doi:10.1016/S0378-1119(00)00472-8]

Fricke, W.F., Seedorf, H., Henne, A., Krüer, M., Liesegang, H., Hedderich, R., Gottschalk, G., Thauer, R.K., 2006. The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H_2 for methane formation and ATP synthesis. *J. Bacteriol*., **188**(2):642-658. [doi:10.1128/JB.188.2.642-658.2006]

Hirokawa, T., Boon-Chieng, S., Mitaku, S., 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**(4):378-379. [doi:10.1093/bioinformatics/14.4.378]

Karkas, J.D., Runder, R., Chargaff, E., 1968. Separation of *B. subtilis* DNA into complementary strands, II. Template functions and composition as determined by transcription with RNA polymerase. *PNAS*, **60**(3):915-920. [doi:10.1073/pnas.60.3.915]

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**(7):283-290.

Karlin, S., Mrázek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol*., **179**(12):3899-3913.

Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ito, T., Ichiyoshi, N., Nishikawa, K., 2002. GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res*., **30**(1):294-298. [doi:10.1093/nar/30.1.294]

Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., *et al*., 2001. Complete genome sequence

of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. *DNA Res*., **8**(4):123-140. [doi:10.1093/dnares/8.4.123]

Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., *et al*., 2000. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. *PNAS*, **97**(26):14257-14262. [doi:10.1073/pnas.97.26.14257]

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., *et al.*, 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, **390**(6658):364-370.

Kreil, D.P., Ouzounis, C.A., 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.*, **29**(7):1608-1615. [doi:10.1093/nar/29.7.1608]

Kumar, S., Tsai, C.J., Nussinov, R., 2000. Factors enhancing protein thermostability. *Protein Eng.*, **13**(3):179-191. [doi:10.1093/protein/13.3.179]

Lambros, R.J., Mortimer, J.R., Forsdyke, D.R., 2003. Optimum growth temperature and the base composition of open reading frames in prokaryotes. *Extremophiles*, **7**(6):443-450. [doi:10.1007/s00792-003-0353-4]

Lawrence, J.G., Ochman, H., 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, **44**(4):383-397. [doi:10.1007/PL00006158]

Lynn, D.J., Singer, G.A.C., Hickey, D.A., 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.*, **30**(19):4272-4277. [doi:10.1093/nar/gkf546]

Mitchell, D., Bridge, R., 2006. A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun*., **340**(1):90-94. [doi:10.1016/j.bbrc.2005.11.160]

Muto, A., Osawa, S., 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *PNAS*, **84**(1):166-169. [doi:10.1073/pnas.84.1.166]

Nakashima, H., Ota, M., Nishikawa, K., Ooi, T., 1998. Gene from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res.*, **5**(5):251-259. [doi:10.1093/dnares/5.5.251]

Nakashima, H., Fukuchi, S., Nishikawa, K., 2003. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J. Biochem.*, **133**(4):507-513. [doi:10.1093/jb/mvg067]

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., *et al*., 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**(6734):323-329.

Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D., Lasky, S.R., Baliga, N., Thorsson, V., Sbrogna, J., *et al*., 2000. Genome sequence of *Halobacterium* species NRC-1. *PNAS*, **97**(22):12176-12181. [doi:10.1073/pnas.190337797]

Paz, A., Mester, D., Baca, I., Nevo, E., Korol, A., 2004. Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *PNAS*, **101**(9):2951-2956. [doi:10.1073/pnas.0308594100]

Runder, R., Karkas, J.D., Chargaff, E., 1968. Separation of *B. subtilis* DNA into complementary strands. III. Direct analysis. *PNAS*, **60**(3):921-922.

Slesarev, A.I., Mezhevaya, K.V., Makarova, K.S., Polushin, N.N., Shcherbinina, O.V., Shakhova, V.V., Belova, G.I., Aravind, L., Natale, D.A., Rogozin, I.B., *et al*., 2002. The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *PNAS*, **99**(7):4644-4649. [doi:10.1073/pnas.032671499]

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S.L., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., *et al*., 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**(6799):959-964.

Yokota, K., Satou, K., Ohki, S., 2006. Comparative analysis of protein thermostability: differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins. *Sci. Technol. Adv. Mater*., **7**(3):255-262. [doi:10.1016/j.stam.2006.03.003]

Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I., 2007. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol*., **3**(1):62-72. [doi:10.1371/journal.pcbi.0030005.eor]

Zhou, X.X., Wang, Y.B., Pan, Y.J., Li, W.F., 2008. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*, **34**(1):25-33. [doi:10.1007/s00726-007-0589-x]