Japanese Journal of Health, Fitness and Nutrition, 17(1), 19 − 24 (2012)

Short Communication

Determination of the G+C Content of Genomic DNA by Reversed Phase HPLC

Jun KOBAYASHI*[1, 2], **Masakatsu NOHARA***[3], **Keiichi IKEDA***[4], **Mariko MOCHIZUKI***[1] **and Hiromi AMAO***[3]

* 1：*Department of Veterinary Nursing and Technology, Faculty of Veterinary Medicine, Nippon Veterinary and Life Science University*

* 2：*Department of Environmental Health, National Institute of Public Health*

* 3：*Department of Animal Science, Faculty of Applied Life Science, Nippon Veterinary and Life Science University*

* 4：*Department of Health Science, Juntendo University School of Health and Sports Science*

Abstract

Deoxyribonucleic acid (DNA) contains the instructions for the development and function of living organisms. It is considered that people's physical strength and other characteristics are in part due to genetic differences. DNA analysis technology has become a useful tool for the discovery of new genes related to human illness, to pinpoint the infection route of pathogenic bacteria and virus.

Analysis of the G+C content of DNA by a precise and rapid method is useful for finding new bacteria and pathogenic bacteria, as well as for estimating the potential causes of disease. We found that the sample storage conditions and data analysis have a role in the accuracy of the determination of the G+C content by reversed phase HPLC. A column that can be used without organic solvent was selected. UV detection (270 nm) was performed at 40 ℃. Four mononucleotides (dNMPs) were separated within 15 min. High reproducibility was attained, with less than 1% CV (125 μM dNMPs). We observed that dNMPs decompose easily when kept at room temperature. In this study, the G+C content was calculated by three different methods: 1) using the ratio of dNMPs to equal moles; 2) using calibration curves of dNMPs; 3) using DNA from real microbes, with known composition, as a standard.

Key words：G+C content（G+C 含量）, HPLC（高速液体クロマトグラフィー）, genomic DNA （ゲノム DNA）

Ⅰ. Introduction

Deoxyribonucleic acid (DNA) contains the instructions for the development and function of living organisms. However, DNA arrangement varies even within the same species, and its individual specificity is great. It is known that the risk of illness is increased by errors in the gene sequences, which include chromosome aberrations. Moreover, it is considered that people's physical strength and other characteristics are in part due to genetic differences. These differences are still not completely understood, so there is great interest in research directed to elucidate their role.

DNA analysis technology has progressed remarkably by the development of the PCR method. The DNA sequencing method, which makes use of PCR, was responsible for the success of the Human Genome Project at the end of the 20th century. This method has become a useful tool for the discovery of new genes related to human illness (e.g.: cancer promoter), to pinpoint the infection route of pathogenic bacteria and virus, and to understand the variation mechanism of influenza[1)]. It should be noted that the way microbes gain new pathogenicity （e.g.: E. *coli* O 157） or a new strain of influenza is born, is by gene mutation. Moreover, it is possible that there are animal and plant species that have not yet been discovered, that could constitute "new species"[2-4)]. For the latter case, features such as enzyme activity and appearance need to be reported together with the result of gene analysis. For example, cell size, colony characteristics, optimum culture conditions, physiological and biochemical characteristics, and fatty acid composition should be reported for the classification of bacteria. Even with the advances in DNA sequencing technology, G+C content measurement is often used as a simple and effective technique to classify an organism. Although the genetic information can only be checked by DNA sequencing or a gene expression experiment, it is clear that organisms of the same kind have a very similar G+C content.

We consider that the accuracy and reporting style of the methods for the determination of G+C content are very important. The current methodology shows neither CV nor SD values when reporting a new species. For this reason, when the values change considerably, it is unclear whether the error is acceptable. This generates problems, because the data in a report describing a new species do not allow readers to judge objectively whether the species in question is new or already known. We thus focused on developing an analytical method using reversed phase HPLC and nuclease P1 enzyme. First, the analytical conditions required to calculate the G+C content were investigated. Then, the accuracy of the G+C content obtained was evaluated, and finally, three calculation methods for the determination of the G+C content were compared.

Ⅱ. Experimental

Reagents

2'-Deoxyadenosine-5'-monophosphoric acid（A; Sigma, St. Louis, MO, USA）, thymidine-5'-monophosphate disodium salt hydrate（T; Sigma）, 2'-deoxyguanosine-5'-monophosphate sodium salt hydrate（G; Sigma）, and 2'-deoxycytidine-5'-monophosphate sodium salt（C; Sigma）were used as standard mononucleotides（dNMPs）. Reference solutions were prepared with concentrations of 15.6–500 μM. Sodium phosphate, phosphoric acid, and ammonium acetate (special grade; Wako Pure Chemicals, Osaka, Japan）were used as mobile phase buffers. Acetonitrile（HPLC grade; Nacalai Tesque, Kyoto, Japan）was used as organic solvent for the elution. Nuclease P1 was obtained from Yamasa Corporation (Chiba, Japan). Four dummy samples with varying compositions were used. Coliform DNA（50% of the theoretical G+C content[5]）was used as a real sample. First, the RNase process was carried out. Then, the process using phenol/chloroform was performed, followed by decomposition using nuclease P1. All other reagents were obtained from commercial sources, and were of the highest quality.

Analytical conditions

The HPLC equipment used consisted of a pump（L-2130; Hitachi, Ibaraki, Japan), a UV-VIS spectrophotometric detector（L-2420; Hitachi), a recorder（D-2500; Hitachi）, and a manual injector with a 20 μL sample loop（7725i; Rheodyne, Oak Harbor, WA, USA）. A YMC Pack ODS AQ reversed-phase column（150 × 6.0 mm i.d.; YMC Co., Kyoto, Japan）was used through the study. The optimized analytical conditions were as follows: the eluent was 10 mM phosphoric acid（pH 2.4）/ 0.5% acetonitrile, the column temperature was set to 40 ℃, the flow rate was 1.0 mL/min, and the detection wavelength was 270 nm. Calculation was performed by the peak area method.

Determination of the G+C content

In this study, the G+C content was calculated by three methods（A-C）. Method A determinates the G+C content (mol %）from the ratio of the peak areas（[G+C]/[A+T+G+C]）, using the dNMPs of the reference solution. Method B sets the concentration of each dNMP based on the absolute calibration curve, and calculates the sum of the concentrations of G and C. Method C expresses the G+C content as a ratio of the result of the analysis of coliform. For this method, it is considered that the decomposed coliform materials represent 50% of the G+C content,[5] and it is assumed that all of the original DNA is quantitatively decomposed by nuclease P1.

Ⅲ. Results and Discussion

Selection of analytical column

At a pH of 2 or more（the limit for most columns' operating conditions）, the protons of the phosphate group in dNMP molecules dissociate partially. It is thought that higher pHs result in a more advanced dissociative state. In the analysis by ionic exchange mode, the strong retention of analytes to the column is achieved by the ionic state. However, in reversed phase mode, the ionic state becomes a factor which weakens the retention of analytes to the column[6,7]. Because each dNMP molecule has purine and pyrimidine ring structures, it is considered that retention in reversed phase mode can be achieved at a low pH. In this study, we found that the retention time can be affected by the organic solvent content in a reversed phase system, so we decided to use an ODS column. Although the use of an organic solvent can shorten the retention time, prolonging it is impossible. Since it was considered that the retention of analytes to the column could become very weak, a column that can be used without the application of organic solvent was chosen.

Tamaoka et al[5]. reported that the HPLC resolution improves when mononucleosides（NMPs）obtained by alkali phosphatase processing of the corresponding dNMPs are used. The rates of the phosphatase enzyme reactions, changes in the partial concentration of dNMPs by deproteinization, or possible procedural complications were not evaluated in this study.

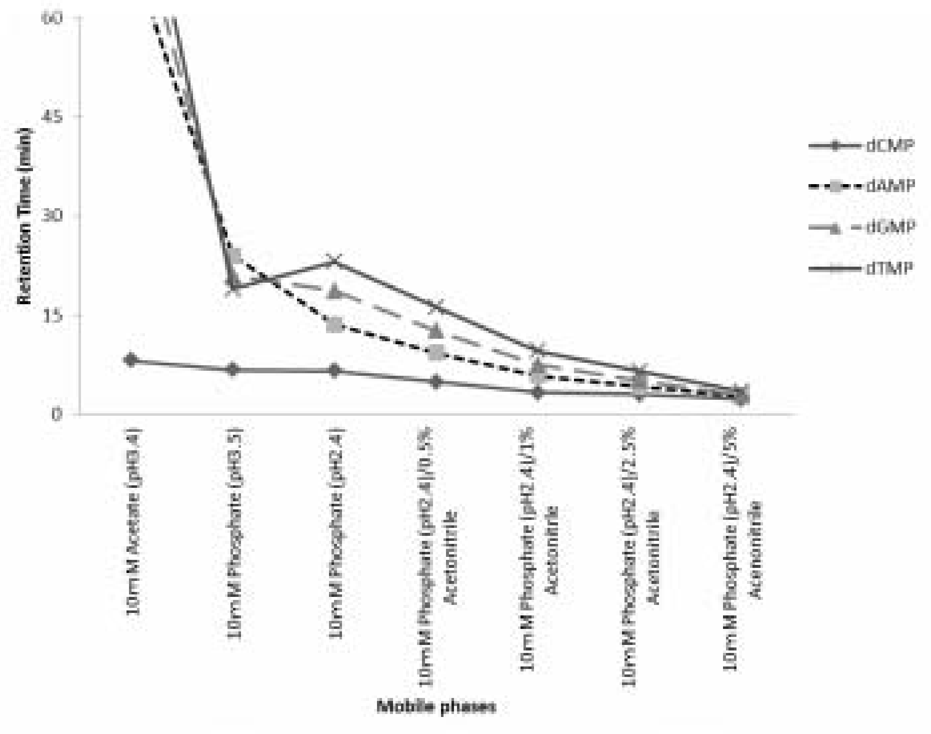

Fig. 1. Resolutions of the four deoxynucleotides

These preliminary experiments were performed at room temperature (22-26 ℃). n = 3.

Determination of analytical conditions

Wavelength changes affect the sensitivity, but for this study this was not an issue since no analysis at trace levels was involved. Because a wavelength of 250−270 nm was used in a previous report[7,8], 270 nm was chosen for this study. Under temperature conditions of 15−35 ℃, 0% of organic solvent, and application of 10 mM phosphate buffer (pH 3.0), the separation of dNMPs in about 30 min was possible, with less than 5% CV (Fig. 1). However, it was judged that a method with high precision and a shorter time was required to calculate the G+C content. Thus, the column temperature was kept at 40℃, using a water bath, and the eluent composition was further examined. Since the retention was too strong when 10 mM ammonium acetate (pH 3.4) was applied, a phenomenon that has been previously reported[8], we decided to use 10 mM phosphoric acid (pH 2.4). Moreover, by adding 0.5% acetonitrile, it became possible to perform analyses with less than 1% CV, within 15 min (125 μM dNMPs).

Quantification by different methods

In this study, we decided to calculate the G+C content with three algorithms. Even when the G+C content was calculated by a molar ratio, it was considered that high linearity between the concentration and the peak area/height was necessary. When linearity was checked using the standard (15.6−500 −M dNMPs mixture), the correlation coefficients between peak area and quality of all dNMPs were 0.999 or more (Table 1) . Minimum determination limits of dAMP, dTMP, dGMP and dCMP were 0.8, 0.2, 0.4 and 1.8 μM, respectively, as determined by the peak area method. Limits of quantitation

Table 1. Figures of merit of the calibration graphs

Analytes	Regression equation[*1]	Correlation coefficient (r^2)	Determination range (μM)[*2]
dAMP	Y = 6041.5 X −12052	0.9999	0.8-500
dTMP	Y = 5846.2 X −29896	0.9998	0.2-500
dGMP	Y = 5569.8 X −20917	0.9997	0.4-500
dCMP	Y = 7148.8 X −17104	0.9998	1.8-500

[*1] Based on peak area. Y, peak area; X, dNMP conc. (μm)

[*2] Calculated by each equation and 10-times SD of 62.5 μM dNMP.

Table 2. Data precision of the proposed method

	dNMPs[*1]	Retention time		Peak area	Peak height
		Average (min)	CV (%)	CV (%)	CV (%)
Within-run (n = 3)	A	5.85	0.00	1.22	0.48
	T	10.23	0.66	0.91	2.53
	G	7.45	0.35	0.35	0.64
	C	3.50	0.14	0.55	0.85
Between-run (n = 3)	A	5.90	0.81	0.54	0.44
	T	10.17	0.96	0.79	2.80
	G	7.41	0.78	6.93	5.97
	C	3.50	0.53	1.60	2.41

[*1] A, dATP; T, dTMP; G, dGMP; C, dCMP; each at a concentration of 500 μM.

obtained by the peak height method were equivalent to those obtained with the peak area method, but had a lower linearity. So, subsequent analyses were performed by the peak area method. For the actual DNA sample, it was ensured that each nucleotide was set to about 250 μM. Results for the actual DNA sample are shown in Table 2. Although for some samples values over 5% CV were seen, the outcome was considered to be acceptable.

Influence of residual RNA and nuclease P1

When actually extracting DNA from bacteria and digesting it with nuclease P1, RNA contamination interferes with the analysis. For this reason, the application of RNase, ethanol precipitation, or phenol/chloroform in advance is thought to be desirable to remove most of the residual RNA. It is thought that nuclease P1 also affects the resulting chromatogram in a similar way. When we tried to digest using a high concentration of enzyme for a short time, enzymatic protein was detected as a broad peak at about 45 min (Fig. 2(B) as an example). If deproteinization is performed independently, it is thought that the concentration of dNMPs changes partially. Therefore, it is not desirable to add this operation. To avoid affecting the analysis, the enzyme activity was considered and it was thought suitable to reduce the quantity of enzyme and to increase the reaction time.

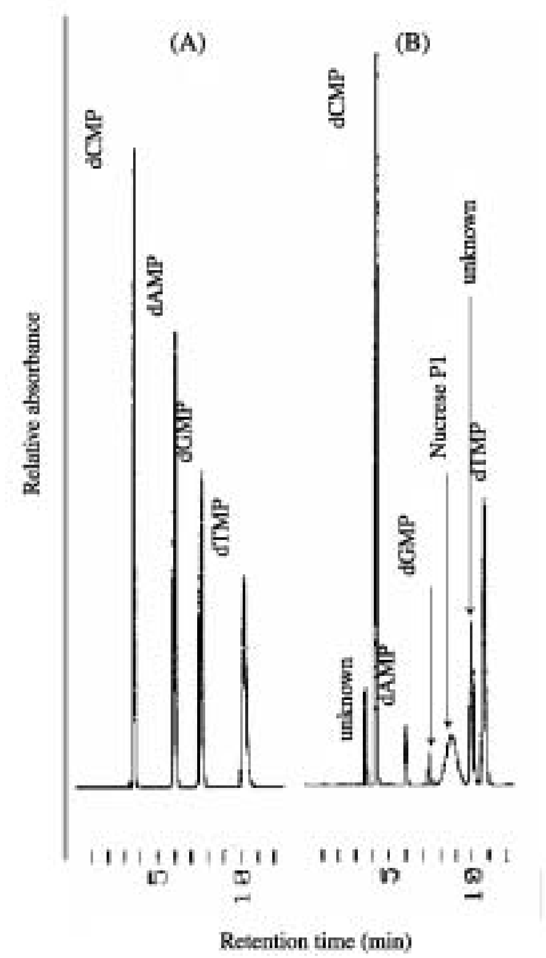

Fig. 2. Degradation of dNMPs over time

(A) Newly prepared standard solution; (B) Two month-old solution (kept at 4℃).

Application of the method in dummy and real samples

This method was applied to dummy and real (coliform digested) samples. The G+C content was calculated by three methods (A-C), and the precision and accuracy were checked. The results are shown in Table 3. For the dummy samples, method A gave results with excellent accuracy and precision. Method B produced the results with the worst precision and accuracy, and an especially large difference of mol% of A and T (the DNA used is not special, so complementarity should be maintained, and A/T and G/C should theoretically have the same mol% values). A gap of several percentage points has been previously reported[9]. However, the value for the real sample obtained by method B was closer to the theoretical value than those obtained with the other methods. The cause for this is not well understood, but it is possible that the measurement error was reduced by two or more data points of the analytical curve. It seems that method B was affected by the purity of the standards used, as well as by instrumental error during weighting. Method A showed good measurement efficiency, and the result was not affected by the kind of sample. Method C was very close to method B. Since other real samples were not measured, a detailed consideration cannot be performed. When coliform DNA was analyzed, very little NMPs of RNA origin were observed in the chromatogram, because digestion by the nuclease was almost perfect (data not shown). Thus, the influence of

Table 3. Data accuracy of the proposed method

Samples(composition)	Method A[*1]				Method B[*1]				Method C[*1]			
	Average	SD	[G-C][*2]	[A-T][*2]	Average	SD	[G-C][*2]	[A-T][*2]	Average	SD	[G-C][*2]	[A-T][*2]
Mock sample-1 (30 mol%)[*3]	30.20	0.00	0.30	0.80	31.94	0.00	2.43	5.31	31.25	0.11	1.87	1.83
Mock sample-2 (70 mol%)[*3]	69.60	0.00	0.34	0.66	66.94	0.01	4.83	1.89	67.14	0.01	1.63	3.94
Mock sample-3 (60 mol%[*3], C:A:G:T=4:3:2:1)	60.00	0.10	–	–	55.98	0.05	–	–	55.80	0.06	–	–
Mock sample-4 (40 mol%[*3], C:A:G:T=1:2:3:4)	39.90	0.10	–	–	38.75	0.05	–	–	38.38	0.04	–	–
Real sample (50 mol%[*3], E. coli)	50.77	0.06	3.34	0.93	49.76	0.07	0.62	4.88	–	–	–	–

[*1] Please see the text for a detailed explanation of the methods
[*2] Absolute value shown in this table
[*3] Theoretical value of the G+C content

residual RNA was negligible. On the other hand, the dNMPs standard solution used in this study was kept refrigerated before analysis. We think that reduction（especially for dAMP and dGMP, as shown in Fig. 2）of the peak by decomposition occurred during analysis, since it was performed at room temperature. It seems necessary to pay close attention to the storage of the reference solution to obtain results with sufficient accuracy; its shelf-life, the use of aliquots, storage below freezing point, or the use of newly prepared solution should be considered. When the DNA of a real sample was stored, a small degree of decomposition was observed, but a detailed examination was not performed.

Proposal for the reporting of data

A previous report[10)] presented data with about 1% error. However, the reproducibility of the method was not indicated. When a method assumes that each DNA nucleotide is responsible for 5% of error in the measurement, the calculation result shows 2.5% of error. From this study, it was considered that a method using LC would generate results with an error of 3–4%. We think that the standard deviation and a detailed description of the method should be included in a report（an example would be “50.3 ± 2.1%, measured by LC”）.

Ⅳ. Conclusions

It is known that the risk of illness is increased by errors in the gene sequences, which include chromosome aberrations. These are still not completely understood, so there is great interest in research directed to elucidate their role. G+C content measurement is often used as a simple and effective technique for DNA analysis.

In this study, the factors affecting the accuracy of the G+C content measurement were examined using a reversed phase HPLC. Recommendations about data reporting were also made. Even when values of G and C were different, the G+C content was determined. However, considering that DNA has complementarity, it is usually thought that the value of G and C is the same. To ensure the reliability of the data, reports should include not only the final result but also information about the accuracy. Until now, most reports using LC have not included the accuracy （CV value） of the data. In this study, we found that about 3–4% of measurement error arises from the LC method used to determine the G+C content. The sequencing method is also not flawless. Naturally, error is also generated by this method. To improve the reliability of the reported values, a description including errors is required.

LC is widely used in many laboratories. Moreover, the usefulness of the G+C content measurement is highly regarded in genetic research, and it is expected that the results and proposal generated by this investigation will be of use.

References

1）Kallioniemi A, Kallionemi OP, Piper J, Tanner M, Stokke T, Chem L, Smith HS, Pinkel D, Gray JW, Waldman FM: Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc Natl Acad Sci USA*, **91**, 2156-2160, 1994

2）Kitahara M, Takamine F, Imamura T, Benno Y: Clostridium hiranonis sp. nov., a human intestinal bacterium with bile acid 7 α-dehydroxylating activity. *Int J Syst Evol Microbiol*, **51**, 39-44, 2001

3) Kitayama M, Sakamoto M, Ike M, Sakata S, Benno Y: Bacteroides plebeius sp. nov. and *Bacteroides coprocola* sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol*, **55**, 2143-2147, 2005

4) Okamoto M, Benno Y, Leung KP, Maeda N: *Bifidobacterium tsurumiense* sp. nov., from hamster dental plaque. *Int J Syst Evol Microbiol*, **58**, 144-148, 2008

5) Tamaoka J, Komagata K: Determination of DNA base composition by reversed-phase high-performance liquid chromatography. *FEMS Microbiol Lett*, **25,** 125-128, 1984

6) Katoh K, Onozaki AS, Ohta T, Ebine H, Kumagai M, Fujimoto M, Kuninaka A: Microbiological identification of single cell proteins based on DNA-GC contents. *Rept. Natl. Food Res Inst*, **43**, 79-89, 1983

7) Kaneko T, Katoh K, Fujimoto M, Kumagai M, Tamaoka J, Katayama-Fujimura Y: Determination of the nucleotide composition of a deoxyribonucleic acid by high-performance liquid chromatography of its enzymatic hydrolysate: a review. *J Microbiol Methods*, **4**, 229-240, 1986

8) Yamanaka S, Kawanishi Y, Okui K: High pressure liquid chromatography analysis of the nucleic acid components in sake. *Hakko kogaku*, **58**, 203-207, 1980

9) Noguchi T, Kumagai M, Kuninaka A: Analysis of base composition of sequenced DNA's by high performance liquid chromatography of their nuclease P1 hydrolysate. *Agric Biol Chem*, **52**, 2355-2356, 1988

10) Agriculture, Forestry and Fisheries Research Council, "The pace of a rice genome base sequence decipherment - the end of a full decipherment-", http://www.s.affrc.go.jp/docs/kankoubutu/ine_genome/pdf/ine_genome2.pdf