Chem-Bio Informatics Journal, Vol. 5, No. 3, pp.56-64(2005)

Comparative proteomics of the prokaryota using secretory proteins

Masahiro Gomi[1], Ryusuke Sawada[2], Masashi Sonoyama[2] and Shigeki Mitaku[2,*]

1 Tokyo University of Agriculture and Technology, Department of Biotechnology, Nakacho 2-24-16, Koganei, Tokyo 184-8588, JAPAN.
2 Nagoya University, School of Engineering, Department of Applied Physics, Furocho, Chikusa-ku, Nagoya 464-8603, JAPAN.

**E-mail: mitaku@nuap.nagoya-u.ac.jp*

(Received October 12, 2005; accepted January 24, 2006; published online January 31, 2006)

Abstract

Secretory proteins function as agents for numerous cell-cell interactions and determine the survival strategies adopted by organisms. Using the SOSUI system for membrane proteins (Hirokawa et al., *Bioinformatics*, 1998) and SOSUIsignal for signal peptides (Gomi et al., *CBIJ*, 2004), we undertook predictive analyses of secretory proteins from 248 prokaryota using all of the amino acid sequences coded by their respective genomes. The number of secretory proteins exhibited a strong positive correlation with the number of total open reading frames, with analysis of these correlations revealing that prokaryotic organisms could be placed into several groups. Symbiotic or obligate parasitic organisms in eukaryotic cells with less than 1200 open reading frames exhibited a single linear relationship between the number of secretory proteins and the total number of open reading frames. Conversely, free-living organisms with more than 2500 open reading frames could be grouped into three linear relationships. The intercept with the axis of the number of open reading frames in the linear relationships was approximately 300 genes for the survival of symbiotic or obligate parasitic organisms and approximately 700 for the free-living organisms. The factor responsible for distinguishing between the different categories of organisms appeared to be G+C content and the number of open reading frames. The roles of secretory proteins and membrane proteins were discussed on the basis of the ratios of those proteins. The list of all predicted secretory proteins for 248 prokaryota is available through the internet at the URL: http://bp.nuap.nagoya-u.ac.jp/sosui/sosuisignal/SOSUIsignalDB/.

Key Words: signal peptide, secretory protein, prokaryote, proteomics, bioinformatics

Area of Interest: Bioinformatics and Bio computing

1. Introduction

The numerous specialized functions of secretory proteins in cells include cell-cell communication, defense mechanisms targeting foreign cells and other related functions. Consequently, the variety of secretory proteins in a proteome is closely related to the survival strategy adopted by a particular organism. A secretory protein has a signal peptide responsible for the translocation of polypeptide segments through the cytoplasmic membrane, but that does not exhibit any real transmembrane segment in the translocated portion. We previously developed software systems, SOSUI [1] and SOSUIsignal [2][3], for predicting membrane proteins and signal peptides. These predictive tools principally employ the physical properties of amino acid sequences and do not rely on sequence homology with known proteins [4][5]. It is therefore expected that the accuracy of the SOSUI and SOSUIsignal predictive tools for unknown sequences from genome sequence analysis is comparable the accuracy associated with known sequences. Furthermore, accuracies of 95% for SOSUI and approximately 90% for SOSUIsignal are considered sufficient for statistical analyses of secretory proteins, which is important for discussions involving the survival strategies adopted by biological organisms.

The large diversity of prokaryotes is likely to have arisen in response to the evolutionary pressure of natural selection and the result of the evolution is stored in their genomes. Obligate parasitic or symbiotic bacteria in eukaryotic cells have smaller genomes than free-living bacteria. The number of open reading frames in the genomes of obligate parasitic or symbiotic bacteria usually number less than 1000, while those of free-living bacteria have more than 1500 open reading frames. Another well-established criterion employed to classify bacteria is Gram staining. Gram negative bacteria are characterized by having an outer membrane that is absent in the Gram positive bacteria and the proteins of their outer membranes and intermembrane spaces are secreted through the innermembrane [6]. In addition, in the three-domain system of biological organisms, the prokaryotes are divided into two kingdoms, the eubacteria and archaea, both of which differ significantly at the genetic level [7][8]. We therefore sought to elucidate the relationship between the ratio of secretory proteins and the survival strategies adopted by organisms in this study. Given that secretory proteins are employed in intercellular communication and defense, comparisons of the proportion of secretory proteins can be a good parameter for the investigation of the survival strategies of organisms.

In the current study, we analyzed all of the amino acid sequences from the genomes of prokaryotes using SOSUIsignal and SOSUI to obtained extensive data sets of secretory proteins. The scatter plot of the number of secretory proteins versus the total number of open reading frames showed that prokaryotes can be classified into several groups that are characterized by the G+C contents and the number of the open reading frames.

2. Methods

2.1 Datasets for entire amino acid sequences in the prokaryota

The amino acid sequences reflecting the genomes of 248 prokaryotic organisms were obtained from the public databases of the NCBI [9]. The organisms analyzed in this study included 22 archaea (A), 20 Gram-positive bacteria with high G+C contents (B1), 63 Gram-positive bacteria with low G+C contents (B2), 105 proteobacteria (B3) and 38 Gram-negative bacteria from various other classes (Bn). Classification of these bacteria used Bergey's classification system [10].

Two genome parameters showed good correlation with the proportion of secretory proteins: the number of open reading frames and G+C content. We calculated the G+C contents for the coding regions of the DNA sequences in the bacterial genomes.

2.2 Software used to predict secretory proteins

Secretory proteins have two basic properties: (1) They have signal peptides that penetrate into the membrane before being cleaved once the translocation process is finished. (2) Other than signal peptides, they do not possess transmembrane segments. The latter attribute, the absence of transmembrane segments, was analyzed using the SOSUI membrane predictor which has been demonstrated to have an accuracy of greater than 95% [1]. This method is based on three physicochemical parameters: the distribution of hydrophobicity, the amphiphilicity of transmembrane helices, and the length of the amino acid sequence of an entire protein [4][5]. The algorithm for the system is very simple but the performance is sufficient for the statistical analysis of genomes. The details of the method have been described previously [1][4][5].

The existence of signal peptides was determined using the predictive SOSUIsignal tool which has an accuracy of approximately 90% [2][3]. Signal peptides have a tripartite structure of amino terminal segments [11], and SOSUIsignal is constituted of three modules which correspond to this tripartite structure. We used the propensities of the occurrence of amino acids for the signal sequences in addition to the same physicochemical parameters as the membrane protein predictor. The details of the method have also been described elsewhere [3].

3. Results and Discussion

We analyzed amino acid sequences derived from genome information of 248 prokaryota and predicted all of the possible membrane and secretory proteins. Figure 1A is a scatter diagram of the 248 organisms showing the number of membrane proteins versus the total number of open reading frames. The corresponding diagram for secretory proteins is shown in Figure 1B. The scatter diagrams for membrane proteins and secretory proteins showed positive correlation between the numbers of predicted proteins and total open reading frames. However, a marked difference was observed in the degree of the scattering of the data points in the two diagrams. The membrane protein data that was well approximated by a straight line crossed the origin:

$$N_{\mathrm{MP}} = 0.22 N_{\mathrm{ORF}} \qquad (1)$$

where N_{MP} and N_{ORF} represent the number of membrane proteins and total number of open reading frames, respectively (Figure 1A) [12]. The r-value of the correlation was as good as 0.95. This equation indicates that the proportion of membrane proteins in the total open reading frames is almost constant at approximately 22%. This high degree of proportionality is probably due to some form of evolutionary selection that has resulted in the maintenance of a constant membrane protein ratio.

Conversely, the corresponding plot for secretory proteins is more diverse. In the diagrams of Figures 1A and 1B, organisms were classified into five groups [10]: archaea (A), Gram-positive bacteria with a high-G+C content (B1), Gram-positive bacteria with low-G+C content (B2), the proteobacteria class (B3) and the remaining Gram-negative bacteria (Bn). Proteobacteria, for example, had approximately twice the number of secretory proteins than the Gram-positive bacteria with low G+C contents. However, the organisms in all categories were observed to converge into a single grouping at approximately 1200 open reading frames and below.

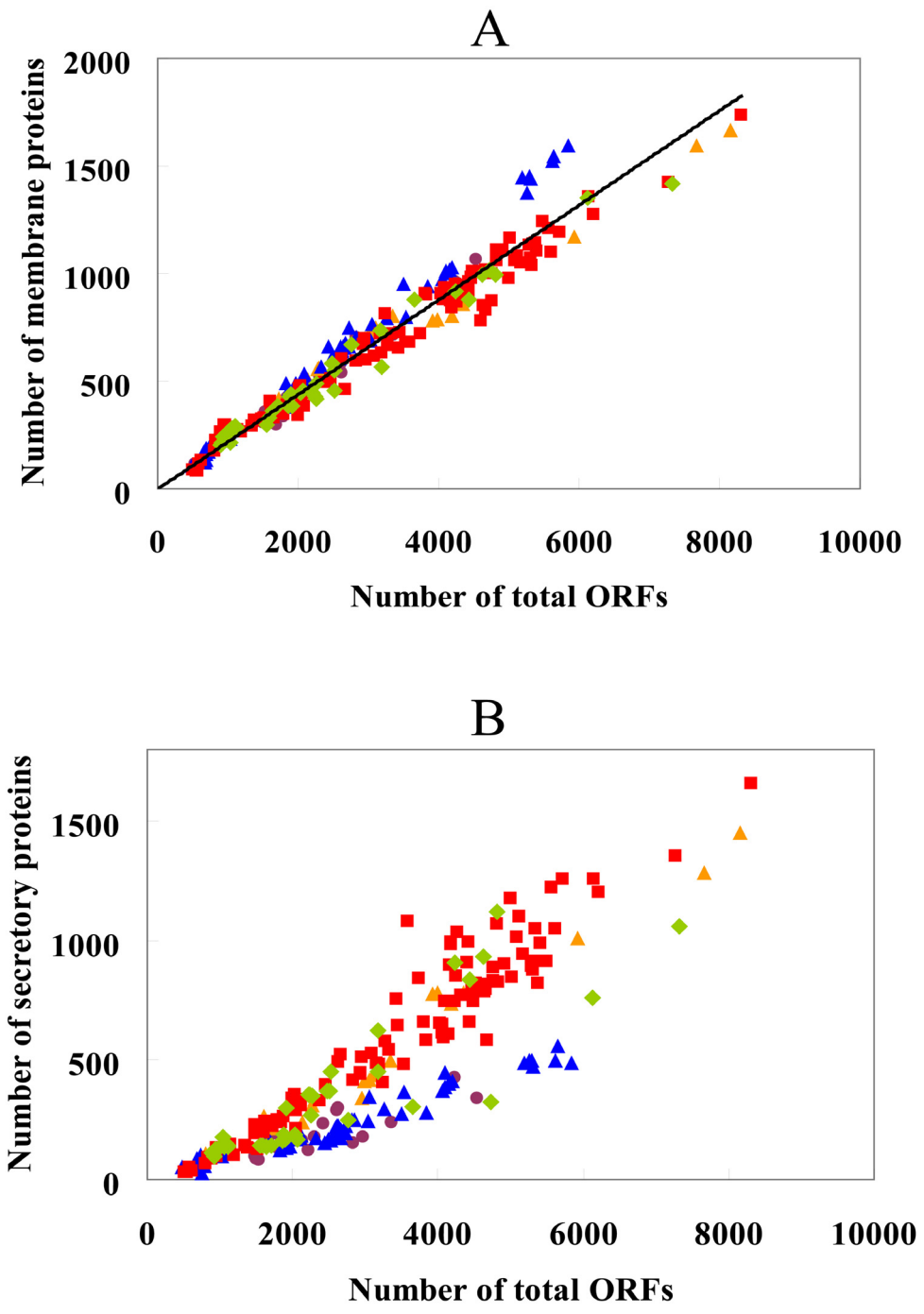

Figure 1. The number of membrane proteins (A) and the secretory proteins (B) plotted as functions of open reading frame numbers for 248 prokaryotes.
The organisms are classified into five categories according to the Bergey's classification method: Archaea (●); Gram positive bacteria with high G+C contents (▲) and low G+C contents (▲); proteobacteria (■) and all other Gram negative bacteria (◆). The scatter diagram for the secretory proteins showed greater scattering of data than that observed for intrinsic membrane proteins. The solid line in the diagram line for membrane proteins shows the best fit for data points using the least square deviation method.

The most significant difference between the membrane proteins and the secretory ones is the intercept with the open reading frame axis. The correlation in Figure 1A for membrane proteins crosses the origin, whereas the intercept for the secretory proteins in Figure 1B is finite. This fact means that the number of the secretory proteins can be very small for an organism to live. On the other hand, membrane proteins are essential components without which a cell cannot live.

The number of the secretory proteins in Gram-negative bacteria was statistically larger than that in Gram-positive bacteria. The secretory proteins in Gram-negative bacteria are located in either of three regions: the outside of a cell, the outer membrane and the intermembrane space. Therefore, it seems reasonable that more proteins are secreted through cytoplasmic membrane in Gram negative bacteria than Gram positive ones which do not have the outer membrane. However, the scattering of data is very large for both groups of Gram positive and negative bacteria.

The scatter diagram of Figure 1B suggests that G+C content and the number of open reading frames in a genome are considerably important factors that influence the ratio of secretory proteins. Therefore, we analyzed the correlation between the ratio of secretory proteins and the total number of open reading frames in the four subgroups of organisms. As shown in Figure 2A, organisms with less than 1200 open reading frames (subgroup I) lie on a line which is represented by the equation:

$$N_{SP} = 0.176(N_{ORF} - 330) \quad (2)$$

The shape of the line from equation (2) for secretory proteins differs from that described by equation (1) of membrane proteins, in that the intercept with the axis describing total open reading frame numbers has a finite value of approximately 300. This means that organisms with a low number of secretory proteins can survive if approximately 300 of the most fundamental genes are present in the genome. In Figures 2B-2D, we showed the scatter plots depicting the ratio of secretory proteins versus the total number of open reading frames for the organisms of the subgroups IIA-IIC that have more than 2500 open reading frames. We omitted the data for organisms with the open reading frame whose number is between 1200 and 2500, because this region is probably the crossover region between the obligate parasite and the free-living organisms. The three subgroups IIA-IIC have different G+C contents; subgroup IIA has G+C contents of less than 0.4, subgroup IIB has G+C contents between 0.4 and 0.55, subgroup IIC has G+C contents greater than 0.55. The relationship between the number of secretory proteins and the total number of open reading frames could be well fitted by the following equations (3), (4) and (5) for subgroups IIA, IIB and IIC, respectively:

$$N_{SP} = 0.102(N_{ORF} - 708) \quad (3)$$

$$N_{SP} = 0.187(N_{ORF} - 747) \quad (4)$$

$$N_{SP} = 0.234(N_{ORF} - 792) \quad (5)$$

The intercept with the open reading frame number axis for organisms in subgroups IIA, IIB and IIC were 708, 747 and 792, respectively. It therefore appears that the intercepts in Figure 2 for free-living organisms are almost constant at approximately 700. Conversely, the proportional coefficients of the linear equations (3)-(5) changed markedly with respect to the G+C content.

The r-values of the correlation were higher than 0.75 except for Figure 2C. The errors of the values of the intercept were estimated by the cross validation test. Randomly taking 70 % of the data, the intercept were calculated by the least square deviation method for 100 times. The scattering of the intercept could be well fitted to the Gauss distribution and the standard deviation were 49, 135, 480 and 209 residues for the groups I, IIA, IIB and IIC, respectively. Except for the group of IIB, in which the scattering of data is large, the intercept for the group I is clearly different from that for the groups II. Therefore, it can concluded that the two parameters, the total number of open reading frames and the G+C contents, are good parameters for classifying prokaryotic organisms.

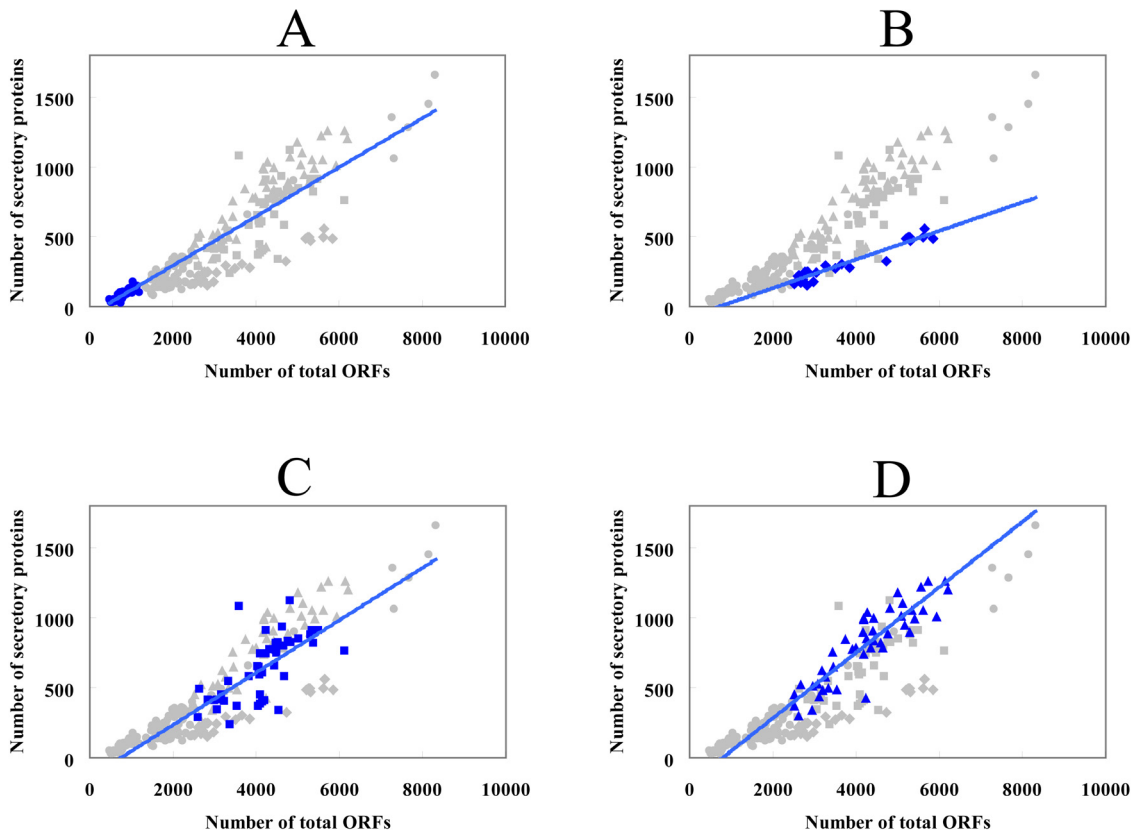

Figure 2. The correlation between numbers of secretory proteins and open reading frames is shown for four groups of prokaryotes with different G+C contents and numbers of open reading frames.

Dark points represent the data points of each selected group: A, group I; B, group IIA: C, group IIB and D, group IIC. The number of open reading frames for organisms in the group I are less than 1200, while those for groups IIA, IIB and IIC range from 2500 to 6500. The ranges in G+C content (R_{GC}) for the three groups are as follows: $R_{GC} < 0.4$ (group IIA), $0.4 < R_{GC} \leq 0.55$ (group IIB) and $R_{GC} \geq 0.55$ (group IIC). Intercepts with the open reading frame axis were approximately 300 for group I and approximately 700 for the groups IIA, IIB and IIC. Organisms with a crossover region between 1200 and 2500 were omitted from the correlation analysis.

Figures 3A and 3B show the two parameters characterizing the linear relationship for the four subgroups of organisms: the intercept with the axis of the open reading frame number (Figure 3A) and the proportional coefficients for the linearity between the numbers of secretory proteins and the total numbers of open reading frames (Figure 3B). The results present very interesting problems on the survival strategies of prokaryotes. The first problem concerns the number of genes for survival. The functions of secretory proteins are related to the types of interactions between the cell and their respective environments. The present results indicate that the number of open reading frames for free-living organisms with very small number of secretory proteins is more than twice that of obligate parasites and symbionts. This finding appears reasonable given that obligate parasites and symbionts acquire many basic nutrients from their host cells. The number of genes at the intercept was about 300 for obligate parasites and symbionts (Figures 3A) which is similar to that of the essential genes of 250 for E. coli [13]. Investigations of which kinds of proteins are necessary for obligate parasites, symbionts and free-living organisms would constitute very interesting material for future research.

The most important point of this work is the accuracy of the prediction for total genomes. The ratio of the number of the unannotated sequences to that of the total ORFs was 37 % in the average of 248 prokaryota, and the corresponding ratio for the subset of the predicted secretory proteins was almost the same level of 36 %. However, the annotation of many ORFs includes the word "putative", "hypothetical", "probable" and "possible". Furthermore, there are ORFs with the annotation, which include "extracelluar", "secretion", "outer membrane", "periplasm". The number of those ORFs in the subset of the predicted secretory proteins was as large as 56%. The remaining ORFs contain in fact scretory proteins whose annotation does not imply the secretion, for example the binding proteins located in the periplasmic space. The false positive data actually exist in the database of the predicted secretory proteins as expected from the error of the prediction systems. The individual methods used in the current work, SOSUI and SOSUIsignal, are accurate enough by cross validation tests [1][3]. Anyway the accuracy of the prediction for the unannotated sequences is crucial for obtaining decisive conclusion from the present results. We show the list of all predicted secretory proteins with the current annotation, SOSUIDBsignal, on the internet at the URL: **http://bp.nuap.nagoya-u.ac.jp/sosui/sosuisignal/SOSUIsignalDB/**.

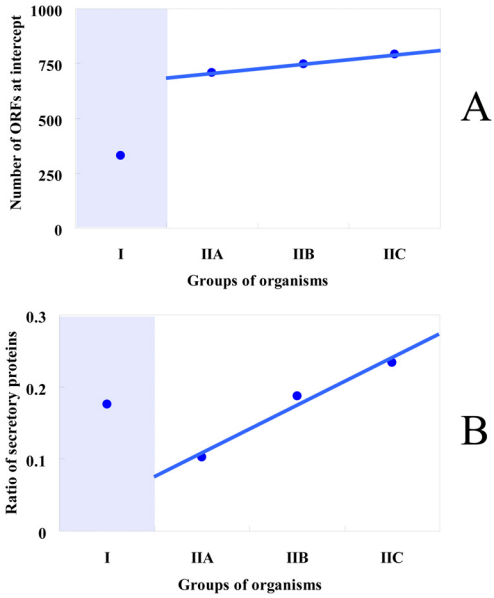

Figure 3. Parameters that affect the correlation between the number of secretory proteins and total open reading frames are shown for four groups of prokaryotes in Figure 2: the intercept with the axis of the open reading frame number for the four groups (A), and the ratio of secretory proteins that correspond to the slope of the linear relationship in Figure 2 (B).

The systematic changes in the coefficient of the linear relationship against G+C content are the second problem. The strong correlation between the proportional coefficients and the G+C contents depicted in Figure 3B probably arose as a consequence of the variation in codon usage in genomes. Therefore, we analyzed the differing propensities in the occurrence of amino acids in total proteomes and plotted these as a function of the G+C contents of the total genomes of the organisms. The results of the analysis are shown in Figures 4A-4C. The amino acids could be classified into ten classes according to their physicochemical properties: alanine, proline, glycine, hydrophobic residues (leucine, isoleucine, valine, phenylalanine, methionine), positively charged residues (lysine, arginine, histidine), negatively charged residues (aspartic acid, glutamic acid),

asparagine and glutamine, small polar residues (serine, threonine), tyrosine and tryptophan, and cysteine. Surprisingly, all of the scatter plots showed very good correlations. Alanine, proline and glycine showed marked positive correlations with the r values of 0.91, 0.75 and 0.87, respectively. Both hydrophobic residues and charged residues decreased in frequency in response to an increase in the G+C contents. These findings suggest that either or all of the three amino acids, alanine, proline and glycine, are responsible for the positive correlation between the ratio of secretory proteins and respective G+C contents. It is likely that the ratio of secretory proteins increases in response to increases in the occurrence of alanine, proline and glycine. In fact, alanine is a strong factor for the signal peptides in the algorithm of SOSUIsignal [2][3]. Similar amino acid bias was reported by Singer and Hickey by the analysis of about 20 set of genome data [14]. The present work showed that the genome dataset ten times larger than the previous work also showed the very good correlation of the amino acid bias. Furthermore, we analyzed only the regions of secretory proteins in genomes, leading to the very similar amino acid biases to those for the corresponding total genomes (Figure 4). In conclusion, one of the possible reasons of this amino acid bias is the formation of secretory proteins.

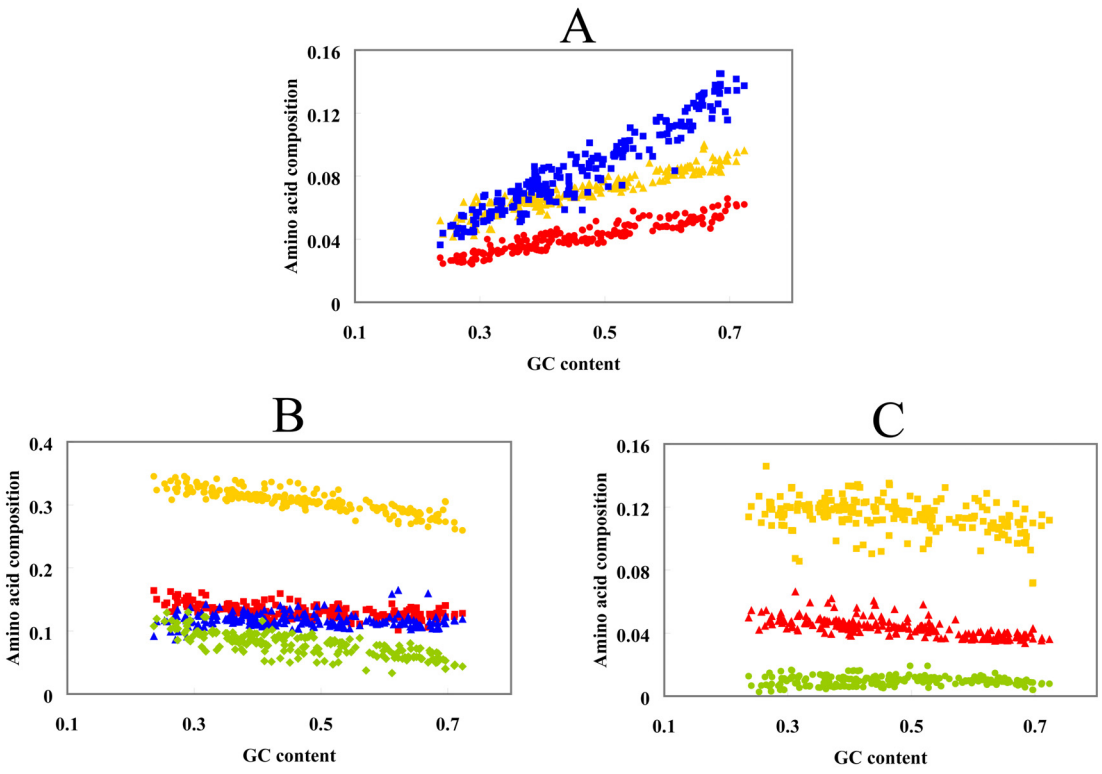

Figure 4. Propensities of the occurrence of amino acids in proteomes are plotted as a function of G+C content.

Amino acids are classified into ten groups according to their physicochemical properties. Amino acids exhibiting a positive correlation were Ala (■), Pro (●) and Gly (▲) (A). Amino acids exhibiting a negative correlation were hydrophobic residues (Leu, Ile, Val, Phe, Met) (●), positively charged residues (Lys, Arg, His) (■), negatively charged residues (Asp, Glu) (▲) and Asn & Gln (◆) (B). Amino acids with almost constant propensity are small polar residues (Ser, Thr) (■), Tyr & Trp (▲) and Cys (●) (C).

Finally, we advocate the use of this physicochemical approach for protein prediction and the understanding of evolutionary relatedness among biological organisms. The physicochemical

approach illustrated here has the potential for predicting the properties of all proteins with the same level of accuracy, including unknown sequences. There is a question whether the accuracy differs depending on the types of organisms, but our previous analyses on the membrane protein prediction did not show large difference in the accuracy. We believe that the current study adequately illustrates the advantages associated with the physicochemical methods used for the prediction of membrane proteins as well as secretory proteins.

This work was partly supported by the Grant-in-Aid for the 21st Century COE "Frontiers of Computational Science" from the Ministry of Education, Culture, Sport, Science and Technology of Japan.

References

[1] T. Hirokawa, B.-C. Seah and S. Mitaku, SOSUI: Classification and secondary structure prediction system for membrane proteins, *Bioinformatics Applications Note*, **14**, 378-379(1998).
[2] M. Gomi, F. Akazawa and S. Mitaku, SOSUIsignal: Software system for prediction of signal peptide and membrane protein, *Genome Informatics*, **11**, 414-415(2000).
[3] M. Gomi, M. Sonoyama and S. Mitaku, High performance system for signal peptide prediction: SOSUIsignal, *Chem-Bio Informatics J*, **4**, 142-147(2004).
[4] S. Mitaku, T. Hirokawa and T. Tsuji, Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces, *Bioinformatics*, **18**, 608-616(2002).
[5] S. Mitaku and T. Hirokawa, Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and protein length, *Protein Engineering*, **12**, 953-957(1999).
[6] M. T. Madigan, J. M. Martinko and J. Parker, Brock Biology of Microorganisms, Eighth ed., Prentice-Hall, Inc., New Jersey: Upper Saddle River, 70-78 (1996).
[7] C. R. Woese, Bacterial Evolution, *Microbiological Reviews*, **51**, 221-271(1987).
[8] C. R. Woese, O. Kandler and M. L. Wheelis, Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. USA*, **87**, 4576-4579(1990).
[9] NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/); RefSeq Release 12.
[10] N. R. Krieg and G. M. Garrity, On Using the Manual. Bergey's Manual of Systematic Bacteriology Second Edition Volume Two The Proteobacteria Part A Introductory Essays, G. M. Garrity, B. J. Brenner, N. R. Krieg and J. T. Krieg eds., Springer Science and Business Media Inc., New York; Springer Street, 15-19 (2005).
[11] K. Nakai, Signal peptides. In Cell Penetrating Peptides, Processes and Applications, CRC Press, 295-323 (2003).
[12] S. Mitaku, M. Ono, T. Hirokawa, B.-C. Seah and M. Sonoyama, Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system, *Biophysical Chemistry*, **82**, 165-171 (1999).
[13] Profiling of E. coli Chromosome (http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp).
[14] G. A. C. Singer and K. A. Hickey, Nucleotide bias causes a genomewide bias in the amino acid composition of proteins, *Mol. Biol. Evol.*, **17**, 1581-1588 (2000).