*J. Biochem.* **143,** 661–665 (2008)
doi:10.1093/jb/mvn017

# Vertebrate Genomes Code Excess Proteins with Charge Periodicity of 28 Residues

**Runcong Ke*, Noriyuki Sakiyama, Ryusuke Sawada, Masashi Sonoyama and Shigeki Mitaku**

*Department of Applied Physics, Graduate School of Engineering, Nagoya University, Nagoya, Japan*

Received December 9, 2007; accepted February 1, 2008; published online February 14, 2008

**All amino acid sequences derived from 248 prokaryotic genomes, 10 invertebrate genomes (plants and fungi) and 10 vertebrate genomes were analysed by the autocorrelation function of charge sequences. The analysis of the total amino acid sequences derived from the 268 biological genomes showed that a significant periodicity of 28 residues is observable for the vertebrate genomes, but not for the other genomes. When proteins with a charge periodicity of 28 residues (PCP28) were selected from the total proteomes, we found that PCP28 in fact exists in all proteomes, but the number of PCP28 is much larger for the vertebrate proteomes than for the other proteomes. Although excess PCP28 in the vertebrate proteomes are only poorly characterized, a detailed inspection of the databases suggests that most excess PCP28 are nuclear proteins.**

**Key words: amino acid sequence, autocorrelation function, charge periodicity, nuclear protein, vertebrate.**

Abbreviations: COG, clusters of orthologous groups; PCP28, proteins with a charge periodicity of 28 residues.

During the evolutionary process, there were several key events, such as the appearance of eukaryotic organisms and vertebrate animals. From the aspect of genomics, these key events are characterized by the appearance of new types of genes that produce new proteins. The computational analysis of the total amino acid sequences from biological genomes is a new tool to study the physical properties of proteins that characterize the great events in evolution. The set of proteins that exist together in a cell form the cellular environment around them. Because cells are crowded with proteins and other molecules, the proteins in a cell must have common physical properties that prevent their aggregation. Furthermore, the new proteins may be translocated to some organelle; for example, a group of translocation factors will go to the nucleus.

We recently identified a group of amino acid sequences coded in the human genome that show a significant charge periodicity of 28 residues. (*1*) This type of protein amounts to ~3% of the total human proteome. Because of the long and very sharp periodicity, proteins with a charge periodicity of 28 residues (PCP28) are anticipated to have related structural and functional characteristics. Although it had been reported that charge clusters in eukaryotic amino acid sequences are related to certain protein functions such as transcriptional factors and developmental control (*2–4*), a detailed analysis of the charge periodicity of amino acid sequences on a genome scale was performed for the first time by our group. We analysed the intra-cellular localization of PCP28 in the human proteome and found that many of these proteins are localized in the nucleus, indicating that they could be transcription factors. However, a previous analysis of amino acid sequences from the *Saccharomyces cerevisiae* genome by the autocorrelation function did not show any peak ascribable to a charge periodicity of 28 residues. (*5*) Therefore, we hypothesize that the ratio of PCP28 in the total ORFs significantly increased during the evolutionary process leading from yeast to humans.

In the present study, we used the autocorrelation function of the charge distribution to analyse all known amino acid sequences from 268 biological genomes, including 248 prokaryotic genomes, 10 invertebrate genomes (plants and fungi) and 10 vertebrate genomes. Our results indicate that the dependence of the number of PCP28 on the total ORFs has two branches: the branch of vertebrates, and the branch of other organisms. The role of PCP28 in the complex biological systems of vertebrates is discussed in the context of the observed charge periodicity.

## MATERIALS AND METHODS

The amino acid sequences of 264 biological genomes were obtained from the NCBI genome database ftp site (ftp://ncbi.nlm.nih.gov/genomes/). The data for *Xenopus tropicalis*, *Fugu rubripes*, *Populus trichocarpa* and *Oryza sativa ssp. japonica* genomes were obtained from the following genome project websites, respectively: http://genome.jgi-psf.org/Xentr4/Xentr4.home.html, http://www.fugu-sg.org/, and http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html and http://rice.tigr.org/. The names

*To whom correspondence should be addressed. Tel: +81 52 788 6218, Fax: +81 52 788 6215, E-mail: ke@bp.nuap.nagoya-u.ac.jp

 

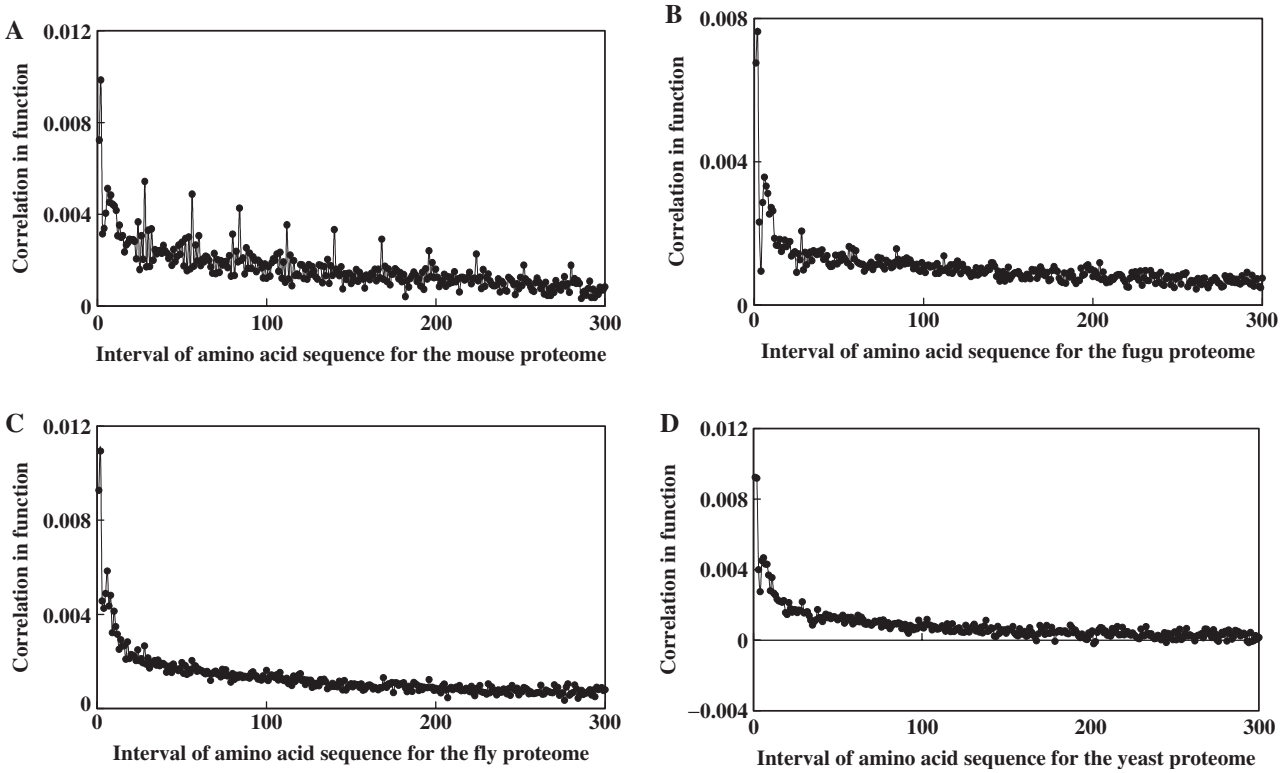


Fig. 1. **Autocorrelation function of all amino acid sequences derived from the vertebrate genomes of mouse (A) and fugu (B) and from the invertebrate genomes of fly (C) and yeast (D).** The plots indicate charge periodicity at intervals of multiples of 28 residues for vertebrates, and no significant periodicity for invertebrates.

of the 268 organisms and their total number of ORFs are available on our web page: http://bp.nuap.nagoya-u.ac.jp/sosui/sosuipcp28/SOSUIDBpcp28.

The autocorrelation function of electric charge C(j) of the sequences was calculated by the following equation (*1*, *5*):

$$C(j) = \frac{\sum_{k=1}^{N} \sum_{i=1}^{L(k)-j} [q_k(i) q_k(i+j)]}{\sum_{k=1}^{N} [L(k) - j]} \quad (1)$$

in which $j$ represents the interval of the correlation, $q_k(i)$ represents the charge of the $i$-th residue in the $k$-th protein (+1 for Lys, Arg and His, −1 for Asp and Glu, 0 for other residues), $L(k)$ represents the length of the $k$-th protein, and $N$ represents the total number of proteins used for the calculation. When we calculate the autocorrelation of a single amino acid sequence, $N$ is set to one.

We selected PCP28 in all amino acid sequences derived from the biological genomes by the following rule (*1*):

$$C(28) \geq \max\{C(i) : 14 \leq i \leq 42, i \neq 28\} + 0.01 \quad (2)$$

This equation shows the relationship between $C(28)$ and the next maximum value of $C(i)$ around $C(28)$: i is between 14 and 42. By using 0.01 as the threshold of the difference, we could accurately discriminate between PCP28 and non-PCP28 in all amino acid sequences.

## RESULTS

We used the autocorrelation function to analyse the charge distribution in all known amino acid sequences coded by 268 genomes. The charge periodicity of 28 residues in the analysis of total proteomes was clearly observable for vertebrates. Figure 1A–D show representative examples of the results for proteomes of the mouse, fugu, fly and yeast, respectively. The autocorrelation function for the mouse proteome showed a very sharp and long-lasting periodicity of 28 residues. The periodicity for the fugu proteome was weak, although the peaks were observable up to the fourth peak. The autocorrelation functions for fly and yeast proteomes did not show any noticeable peaks at 28-residue intervals.

Each PCP28 could be discriminated by Eq. 2; this equation indicates that the value of the autocorrelation function for a PCP28 with an interval of 28 residues was larger, by 0.01, than the corresponding values in the neighbourhood of the interval region between 14 and 42 residues. Figure 2A shows the autocorrelation of PCP28 in the mouse proteome. Since only the PCP28 data were analysed for this diagram, the peaks were much more apparent than in the corresponding analysis of the total proteome (Fig. 1A). It should be noted that the broad positive correlation in Fig. 1A was not found in Fig. 2A, because the broad positive correlation was a property of the proteins lacking the charge periodicity of 28 residues, as shown in Fig. 2B. Figure 2C and D show the

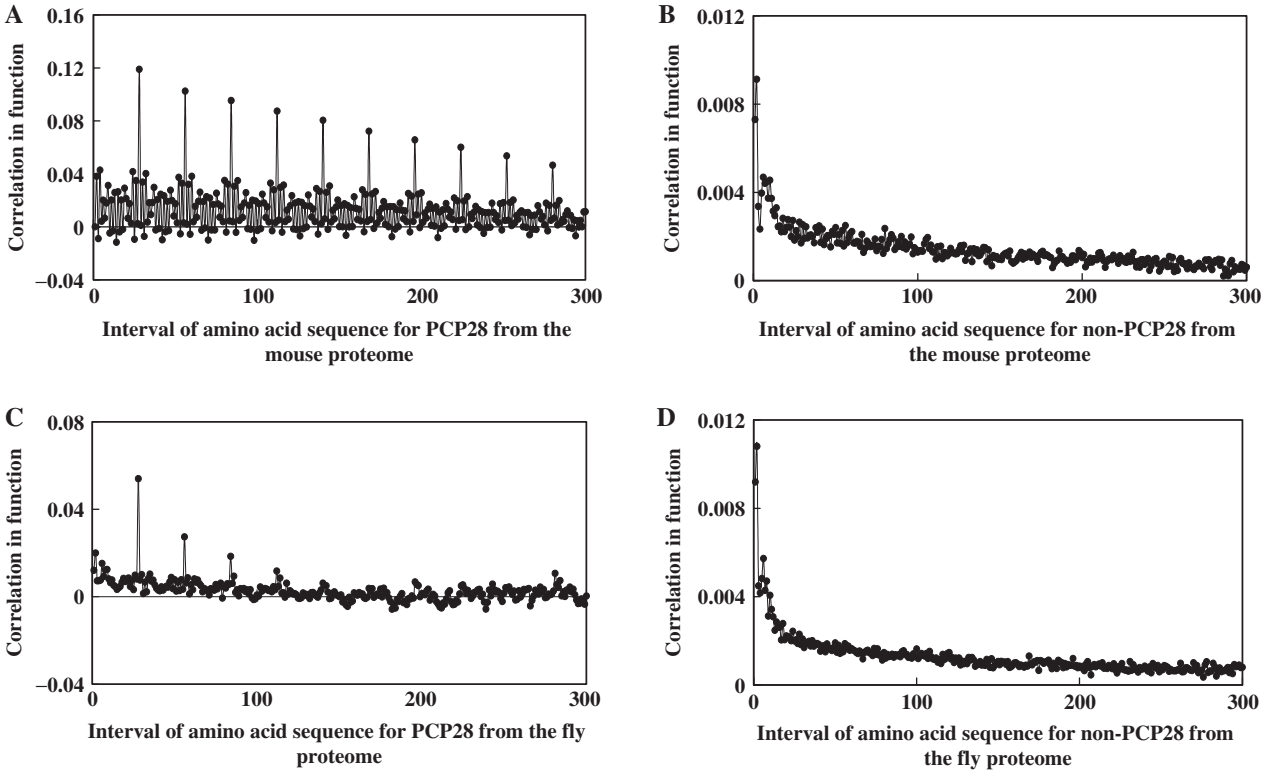


Fig. 2. **The proteins with charge periodicity of 28 residues (PCP28) and the non-PCP28 proteins were separated in the mouse (A and B) and fly (C and D) proteomes.** Significant peaks at intervals of 28 residues could be identified in the PCP28 from the fly genome (C), whereas charge periodicity in the set of all fly amino acid sequences was not observed (Fig. 1C). The autocorrelation function of non-PCP28 with broad positive correlation from the fly proteome (D) was very similar to that from the mouse proteome (B).

autocorrelation of the charge distribution of PCP28 and non-PCP28 in the fly genome, respectively. Our analysis of the total fly proteome (Fig. 1C) did not reveal a charge periodicity of 28 residues. However, after selecting for the PCP28, significant peaks with an interval of 28 residues could be clearly identified, as shown in Fig. 2C. The autocorrelation function for the proteins other than PCP28 in the fly proteome (Fig. 2D) appeared very similar to that of the mouse proteome (Fig. 2B).

We also calculated the independent autocorrelation functions of the positive- and negative charge distributions for amino acid sequences from the human genomes. Both independent autocorrelation functions showed only small peaks at intervals of multiples of 28 residues, indicating that the combination of the positive and negative charges enhanced the charge periodicity of 28 residues (data not shown).

We discriminated all PCP28 from the total proteomes of the 268 biological organisms using the rule described by Eq. 2. The database of the annotations of predicted PCP28 from the 268 genomes is available on the web: Database of Predicted Proteins with the Charge Periodicity of 28 Residues (PCP28): SOSUIDBpcp28 (http://bp.nuap.nagoya-u.ac.jp/sosui/sosuipcp28/SOSUIDBpcp28). The number of PCP28 was significantly correlated to the total number of ORFs for the 10 vertebrate genomes, as well as the 10 invertebrate and 248 prokaryote genomes combined together, as plotted in Fig. 3A. The result clearly indicates that the correlation between the numbers of PCP28 and total ORFs fall into two branches. The vertebrate data constitute the branch with the higher ratio of PCP28 and has a correlation coefficient of 0.95. All other organisms comprise the other branch, in which the correlation coefficient is 0.98. Figure 3B is an enlarged diagram of the prokaryotic proteome data, indicating that the proportionality between the numbers of PCP28 and the total ORFs is almost the same as that for all prokaryota and invertebrates together. The correlation coefficient for prokaryota alone is 0.90. The systematic increase of PCP28 in vertebrates and all other organisms (Fig. 3A) is described by the equations: $y = 0.0079x + 0.0243$ $(x - 12945)$ and $y = 0.0079x$, in which $y$ and $x$ are the numbers of PCP28 and total ORFs, respectively. This dependence suggests that new types of PCP28 emerged as the number of ORFs rose above about 13,000, which is the minimum number of ORFs generally observed in vertebrates.

## DISCUSSION

The two key results of the present study are the proportionality of the number of PCP28 to that of the total ORFs, and the significant increase of this type of protein in vertebrates. The plot of all non-vertebrate organisms correlated reasonably well to a straight line in the dispersion diagram of the numbers of PCP28 *versus* that of the total ORFs (Fig. 3A), indicating that the PCP28 ratio

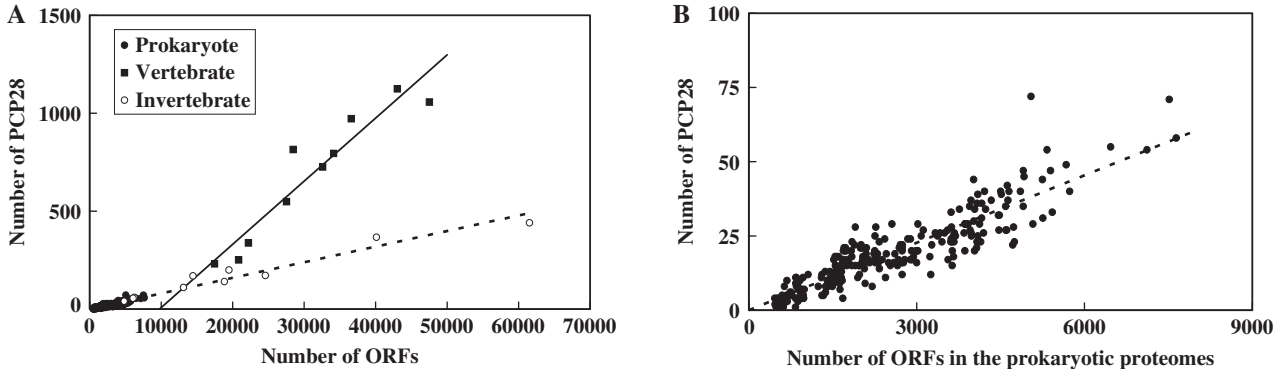


Fig. 3. **Relationship between the number of PCP28 in biological genomes and the number of ORFs (A).** Two branches with significant correlations were observed. One branch was the vertebrates (filled squares), which have a higher ratio of PCP28 and a correlation coefficient of 0.95. The other branch was the set of invertebrates (open circles) and prokaryotes (filled circles) together, for which the correlation coefficient is 0.98. The distribution for prokaryotic proteomes was enlarged (B), indicating that the ratio of the number of PCP28 to the total ORFs in prokaryotic proteomes is almost the same as that in the set of all prokaryotes and invertebrates together. The correlation coefficient for prokaryota alone is 0.90.

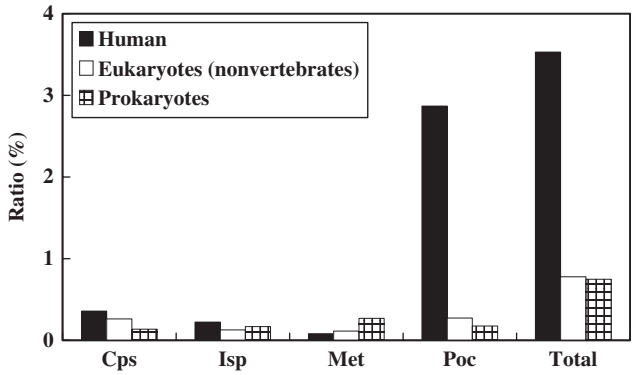


Fig. 4. **Four functional categories for human PCP28 and PCP28 from groups of five non-vertebrate eukaryotes and 64 prokaryotes according to the functional database COG (Clusters of Orthologous Groups of Proteins).** Isp, information storage and processing; Cps, cellular processes and signalling; Met, metabolism; and Poc, poorly characterized. The distribution of functions was similar among the organisms, with the exception that the fraction of poorly characterized human PCP28 was significantly greater than the corresponding fractions for non-vertebrate eukaryotes and prokaryotes.

is fairly constant for these biological organisms. These observations strongly suggest the existence of a mechanism that maintains this ratio, which is closely related to the evolutionary processes.

When the annotations of PCP28 were investigated according to the functional classification database COG (Clusters of Orthologous Groups of Proteins) (http://www.ncbi.nlm.nih.gov/COG) (*6*, *7*), we found that PCP28 have various biological functions. Figure 4 shows the ratios for four functional categories of PCP28: information storage and processing (Isp), cellular processes and signalling (Cps), metabolism (Met) and poorly characterized (Poc). The ratios for three groups of organisms are plotted: prokaryota, non-vertebrate eukaryota and human (as the representative vertebrate). The prokaryota include 13 archaea and 51 eubacteria; the five eukaryota investigated are *S. cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*. Most PCP28 seem to be helix-rich proteins, as seen from the structures of known proteins. However, the variety of biological functions is wide and the distribution of the functions is very similar among invertebrate organisms. The poorly characterized proteins are dramatically increased in the human proteome, indicating that new groups of PCP28 emerged in the vertebrate proteomes (Fig. 4).

We previously examined the intra-cellular localization of PCP28 by analysing the protein database SWISS-PROT. (*8*) Although the intra-cellular localization of PCP28 is only partially known, a large fraction of PCP28 with known localization is in the nucleus. (*1*) Many zinc finger proteins with multiple fingers contact the DNA with a charge periodicity of around 30 residues; some examples are the mouse Zif268–DNA complex with three fingers (*9*), the human GLI–DNA complex with five fingers (*10*) and the *Xenopus laevis* TFIIIA–DNA complex with six fingers (*11*). Therefore, it is reasonable to hypothesize that many of the nuclear PCP28 are likely to be transcription factors. If these PCP28 are in fact the transcription factors that emerged at the dawn of vertebrates, they are very important proteins to be investigated.

In order to study the relationship between zinc finger proteins and PCP28, 718 C2H2-type zinc finger proteins from the human genome were analysed by the auto-correlation functions of charge distribution. The number of PCP28 in C2H2 type zinc finger proteins was 546, and that of non-PCP28 was 172. The number of zinc finger domains in proteins showed very different histograms between PCP28 and non-PCP28, as shown in Fig. 5A. The peak of the number distribution of C2H2 zinc finger domains for PCP28 was eight, whereas that for non-PCP28 was two. Therefore, it seems that the periodicity of 28 residues is simply due to the tandem of zinc finger domains. However, this clear periodicity of 28 residues has not been reported. Furthermore, the autocorrelation analysis for a single type of amino acid did not show any sharp periodicity of 28 residues, and a similar analysis for positive and negative charges did not show significant

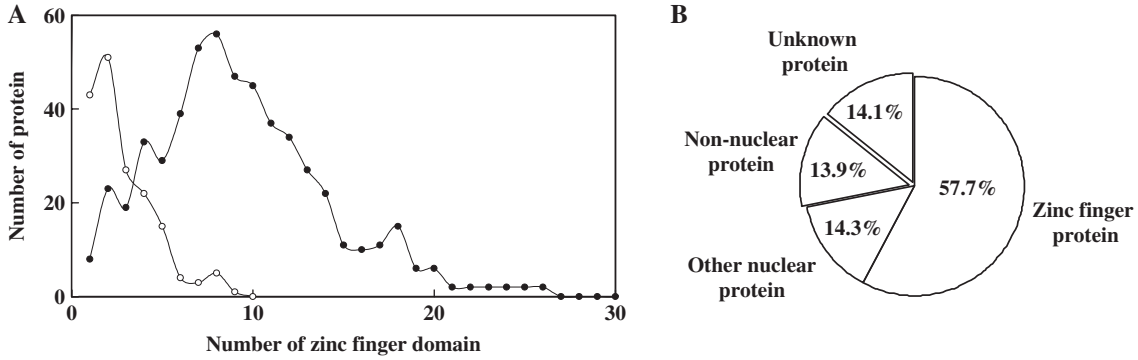


Fig. 5. **The number distributions of C2H2-type domains in PCP28 and non-PCP28 zinc finger proteins from the human genome (A) and the ratio of PCP28 of various functions to all human PCP28 (B).** The number distribution of C2H2 type zinc finger domains was very different between PCP28-type zinc finger proteins (filled circles) and non-PCP28-type zinc finger proteins (open circles). The peaks of the number distribution of C2H2 zinc finger domains were eight for PCP28 and two for non-PCP28 (A). When the nuclear PCP28 were further classified, we found that there are many other types of nuclear proteins aside from zinc finger proteins, indicating that the charge periodicity of 28 residues is a rather universal characteristic of the nuclear proteins (B).

peaks in the autocorrelation functions. Those results indicate that the net charge distribution is one of the essential factors for the formation of this category of proteins.

We also classified 972 PCP28 from the human genome into four categories (Fig. 5B): zinc finger proteins, other nuclear proteins, non-nuclear proteins and unknown proteins. The ratios of zinc finger proteins and other nuclear proteins are 57.7% and 14.3%, respectively. The category of other nuclear proteins included a variety of proteins such as DNA-binding proteins and ribosomal proteins, suggesting that the charge periodicity of 28 residues in amino acid sequence is a universal characteristic for the localization of proteins in the nucleus, and that this property can therefore be used to develop software systems for predicting nuclear proteins.

Several challenges arise from the present results. The first challenge is to develop software for discriminating the excess nuclear PCP28 in vertebrates. We are now investigating the physical properties of PCP28 to enable the development of such software, which will be described elsewhere together with the database of the nuclear PCP28. The second challenge is to determine why PCP28 have this very sharp periodicity of 28 residues. A very large positive correlation in the charge distribution means that amino acid segments within a protein probably interact with each other by electrostatic repulsion. (*5*) The analysis of the dynamic structure of PCP28 will provide additional insights into the reasons for the evolutionary development of this type of protein.

This work was partly supported by the 21st Century COE of Frontiers of Computational Science at Nagoya University.

REFERENCES

1. Ke, R., Sakiyama, N., Sawada, R., Sonoyama, M., and Mitaku, S. (2007) Human genome encodes many proteins with charge periodicity of 28 residues. *Jpn. J. Appl. Phys.* **46**, 6083–6086
2. Brendel, V. and Karlin, S. (1989) Association of charge clusters with functional domains of cellular transcription factors. *Proc. Natl. Acad. Sci. USA* **86**, 5698–5702
3. Karlin, S. (1995) Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.* **5**, 360–371
4. Karlin, S., Mrazek, J., and Gentles, A.J. (2003) Genome comparisons and analysis. *Curr. Opin. Struct. Biol.* **13**, 344–352
5. Ke, R. and Mitaku, S. (2005) Local repulsion in protein structures as revealed by a charge distribution analysis of all amino acid sequences from the *Saccharomyces cerevisiae* genome. *J. Phys. Condens. Matter* **17**, S2825–S2831
6. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997) A genomic perspective on protein families. *Science* **278**, 631–637
7. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28
8. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370
9. Pavletich, N.P. and Pabo, C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809–817
10. Pavletich, N.P. and Pabo, C.O. (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science* **261**, 1701–1707
11. Nolte, R.T., Conlin, R.M., Harrison, S.C., and Brown, R.S. (1998) Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor. *Proc. Natl Acad. Sci. USA* **95**, 2938–2943