

# Analysis of the Proteins with Positive Charge Autocorrelation in Amino Acid Sequences from Eukaryotes

**Runcong Ke**  
ke@bp.nuap.nagoya-u.ac.jp

**Masashi Sonoyama**  
sonoyama@nuap.nagoya-u.ac.jp

**Shigeki Mitaku**  
mitaku@nuap.nagoya-u.ac.jp

Department of Applied Physics, Graduate School of Engineering, Nagoya University, Furocho, Chikusa-ku, Nagoya 464-8603, Japan

**Keywords:** charge autocorrelation, eukaryote, nuclear protein, secretory protein

## 1 Introduction

Charge-charge interactions within protein are important for its protein structure and function. Such electrostatic effects on an individual protein structure are well investigated by experimental and theoretical studies. However, common features of such effects on all proteins encoded by organism genomes have hardly been investigated yet. Previously, we defined the charge autocorrelation function based on the average value of charge-charge pairs in an amino acid sequence and calculated this function of all amino acid sequences in eukaryotes and prokaryotes [2, 3]. The distributions of charge autocorrelation functions showed significant positive correlation for eukaryotes and no correlation for prokaryotes. Furthermore, there is the distribution of charge periodicity of 28 residues in vertebrate genomes [3]. We extracted the proteins with charge periodicity of 28 residues (PCP28) from vertebrate genomes and analyzed the functions of these proteins, finding that many PCP28 belong to the nuclear protein such as DNA binding protein.

In this work, considering that there are the positive correlations in the distributions of charge autocorrelation in all amino acid sequences from eukaryotes, we extracted the proteins with positive charge autocorrelation (PPCA) from eukaryotic genomes. When we investigated known functions of PPCA in eukaryotes, we found that many of these PPCA were the nuclear proteins and secretory proteins.

## 2 Method and Results

The total open reading frames (ORFs) of 20 eukaryotic genomes, 10 vertebrates and 10 invertebrates (including plants and fungus), were mainly obtained from the database NCBI [5]. We selected the PPCA in amino acid sequences, excluding the proteins with charge periodicity of 28 residues (PCP28), from the eukaryotic genomes by the following formulas:

$$C(j) = \frac{\sum_{i=1, j=1}^L [q(i)q(i+j)]}{L-j} \quad (L \geq i+j) \quad (1)$$

$$\sum_{j=6}^{10} C(j)/5 > 0.01 \quad (2)$$

in which  $C(j)$  represents the charge autocorrelation function in amino acid sequences,  $j$  the interval of the autocorrelation functions in sequences,  $q(i)$  the charge of the  $i$ -th residue (+1 for Lys, Arg and His, -1 for

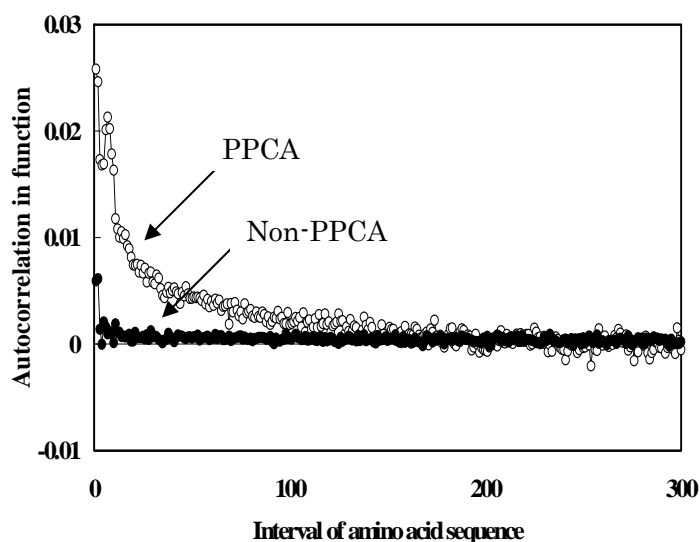


Figure 1: The distribution of charge autocorrelation function from PPCA and non-PPCA in yeast genome. 1283 PPCA were extracted from all amino acid sequences by Eq.(1) and (2). The distribution shows significant positive correlation in PPCA and no correlation in non-PPCA. Many of PPCA from eukaryotic genomes are nuclear proteins and secretory proteins, suggesting that the positive charge autocorrelation in sequences was probably related to the signal peptide.

Asp and Glu, 0 for other residues), and  $L$  the length of the protein. We can accurately discriminate between PPCA and non-PPCA in all amino acid sequences with formula (1, 2).

Figure 1 shows the distributions of charge autocorrelation functions from PPCA and non-PPCA in yeast genome. The average ratio of PPCA to total ORFs from eukaryotes was approximately 20%. The known functions of PPCA in human were investigated according to their intracellular localization in SwissProt. The nuclear protein and secretory protein (including plasma membrane protein) were 45% and 28% to all PPCA, respectively.

### 3 Discussions

In this work, we extracted the proteins with positive charge autocorrelation (PPCA) in all amino acid sequences from eukaryote genomes and most of known PPCA were the nuclear protein and secretory protein (including plasma membrane protein). It has been known that there is signal peptide with rich charged residues in many nuclear protein and secretory protein [1, 4]. Therefore, we considered that the positive charge autocorrelation in amino acid sequence was related to the signal peptide and could be used to prediction of nuclear protein as a parameter.

Because the number of PPCA in eukaryotes was significantly higher than that in prokaryotes (data not shown), the number of PPCA has increased in the process of evolution from prokaryotes to eukaryotes.

### References

- [1] Kalderon D., Richardson W.D., Markham A.F. and Smith A.E., Sequence requirements for nuclear location of simian virus 40 large-T antigen, *Nature* 311: 33-38, 1984.
- [2] Ke R. and Mitaku S., Local repulsion in protein structures as revealed by charge distribution analysis of all amino acid sequences from the *Saccharomyces Cerevisiae* genome, *J. Phys. Condens. Matter.*, 17: S2825-S2831, 2005.
- [3] Ke R., Sakiyama N., Sawada R., Sonoyama M. and Mitaku S., Vertebrate genomes code excess proteins with charge periodicity of 28 residues, *J. Biochem.*, 143(5): 661-665, 2008.
- [4] von Heijne, G., Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells, *EMBO J.*, 3(10): 2315-2318, 1984.
- [5] URL: <ftp://ftp.ncbi.nih.gov/genbank/genomes/>