Local repulsion in protein structures as revealed by a charge distribution analysis of all amino acid sequences from the *Saccharomyces cerevisiae* genome

2005 J. Phys.: Condens. Matter 17 S2825

(http://iopscience.iop.org/0953-8984/17/31/007)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **17** (2005) S2825–S2831 doi:10.1088/0953-8984/17/31/007

Local repulsion in protein structures as revealed by a charge distribution analysis of all amino acid sequences from the *Saccharomyces cerevisiae* genome

Runcong Ke and Shigeki Mitaku

Department of Applied Physics, Nagoya University, Furocho, Chikusa-ku, Nagoya 464-8606, Japan

Received 21 December 2004, in final form 24 January 2005
Published 22 July 2005
Online at stacks.iop.org/JPhysCM/17/S2825

Abstract

The structures and physical properties of individual protein molecules have been extensively studied, but the general features of all proteins in a cell have hardly been investigated. The distribution of net electric charges of all proteins from the *Saccharomyces cerevisiae* proteome agreed well with a Gaussian distribution. The shift in charge distribution caused by protonation of histidine suggested that the proteins in a cell are buffered against pH changes. A comparison between the amino acid sequences from the proteome and randomly generated sequences indicated that electric charges in the real sequences are clustered. Analysis of autocorrelation function of charged residues in the total proteome of *S. cerevisiae* showed a positive correlation of net charges in amino acid sequences with characteristic length as long as 81 residues, leading to the conclusion that the interactions within proteins is repulsive on average.

1. Introduction

Biological materials such as DNA, proteins, and lipid membranes play unique roles in biological organisms. The mechanism of information storage and processing by DNA was revealed about a half century ago upon the discovery of the hydrogen bonding interactions of base pairs in the double helix structure [1]. However, the structures of proteins are quite diverse and complicated, preventing the revelation of general characteristics of the interactions within protein structures [2]. Recently, we found that the distribution of the net charges of all proteins in a proteome from the *Drosophila melanogaster* genome fit well into a Gaussian distribution in which the mean value was close to zero [3, 4]. The discovery indicates that all proteins in a cell behave collectively, leading to a Gaussian distribution of the net charges. This behaviour of a proteome must be the result of evolutionary processes, although the evolutionary pressure on the charge distribution is not understood.

Two questions arise from the Gaussian distribution of net charges in a proteome. What type of evolutionary pressure exerted on proteins in a proteome results in a Gaussian distribution of

Table 1. Proportion of charged residues of all proteins from the *S. cerevisiae* genome.

Amino acid	pK	Elementary charges	Proportion
Arg	10.8	+1	0.1176
Lys	9.7	+1	
His	7.6	0, +0.5, +1	0.0215
Glu	3.2	−1	0.1230
Asp	2.8	−1	
Other residues	—	0	0.7379

net charges? What type of charge correlation in the amino acid sequences realizes the Gaussian distribution?

In this work, we studied the distribution of electric net charges of proteins for *Saccharomyces cerevisiae* and found that the charge distribution was very similar to that of proteins for *D. melanogaster* [3]. The effect of histidine residue protonation on the charge distribution of all the proteins clearly indicated that the protein system in a cell can provide buffering action near neutral pH. The distribution of electric charges in proteins in a proteome of *S. cerevisiae* was studied in greater detail by calculating the autocorrelation function. The significant positive correlation of charges strongly suggests that the repulsive interaction is statistically dominant within proteins.

2. Dataset and methods

All amino acid sequences from the complete genome of *S. cerevisiae* were obtained from NCBI (ftp://ftp.ncbi.nih.gov/genbank/genomes/S_cerevisiae/) [5, 6]. The genome contains 6217 open reading frames (ORFs) [5, 7].

The total net charges q_{sum} of a protein were calculated by the following equation:

$$q_{\mathrm{sum}} = \sum_{i=1}^{L} q(i) \tag{1}$$

in which L is the length of an amino acid sequence; the electric charge $q(i)$ of the ith amino acid was 1, −1, or 0 for positive, negative, or neutral residues, respectively. Only the electric charge of histidine was varied from 0 to 1 for studying the effect of the protonation condition of this amino acid on the shape of the charge distribution. Table 1 shows the ratios of charged residues in proteins of the *S. cerevisiae* proteome. The ratio of positively charged residues, Arg and Lys, is slightly smaller than that of the negatively charged residues Glu and Asp. Approximately 2% of all amino acids are His, which can have different degrees of protonation depending on the pH environment.

The autocorrelation function $C(j)$ of all amino acid sequences was defined by

$$C(j) = \frac{1}{N} \sum_{k=1}^{N} \left\{ \frac{1}{L(k) - j} \sum_{i=1}^{L(k)-j} [q(i)q(i+j)] \right\} \tag{2}$$

in which $q(i)$ is the charge of the ith amino acid, $L(k)$ represents the length of the kth protein, and N is the total number of proteins in a proteome. This equation represents the correlation of electric charges, which averages out the noise in the sequences.

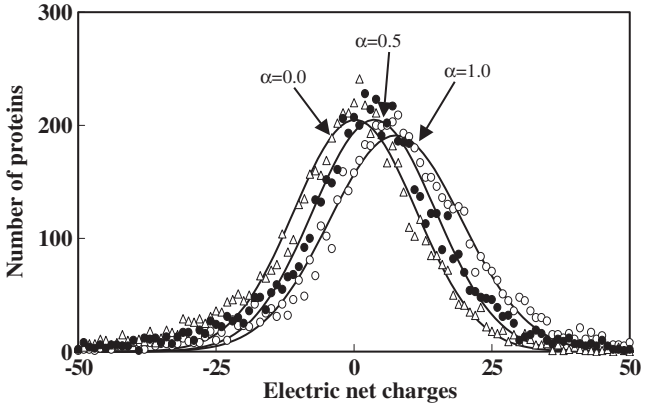

Figure 1. The distribution of net electric charges of all proteins in the proteome of *S. cerevisiae*. Lys and Arg are positively charged and Asp and Glu are negatively charged. The charge of His varied because the pK of His (7.6) is near neutral. The solid curves show the Gaussian distributions best fitted to the data for the degree of protonation α of His, 0, 0.5 and 1.

Table 2. Mean values and standard deviations of charge distributions of all proteins from the *S. cerevisiae* genome for different degrees of dissociation of His.

Degree of dissociation	Mean	Standard deviation
0	0.21	15.7
0.5	3.66	15.8
1.0	7.56	17.0

3. Results

The net charges of all proteins in the *S. cerevisiae* proteome were calculated for three conditions of histidine residue protonation: degree of protonation of 0, 0.5, and 1.0. Figure 1 shows the distributions of proteins as a function of the net charges. The most interesting feature of the distributions is that they fit well with a single Gaussian distribution [3, 4]. The solid curves are the Gaussian distributions in which the mean value $\bar{q}$ and the standard deviation σ are adjusted to give the least square deviation:

$$f(q_{\text{sum}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(q_{\text{sum}} - \bar{q})^2}{2\sigma^2}\right). \tag{3}$$

Table 2 shows the means and standard deviations under different histidine conditions. The standard deviation is almost constant, indicating that the shape of the distribution of net charges is unchanged. Approximately 2% of all amino acids are His, which amounts to about 20% of positively charged residues. Therefore, the shape of the charge distribution is robust against a fairly large change in the protonation of the amino acids.

The mean values of the charge distribution shifted from zero to about seven elementary charges by the protonation of histidine residues. Previously, we analysed the size dependence of the charge distribution for *D. melanogaster*, which indicated that the amino acid sequences have a small number of positive charges on average at the amino terminal end [3]. The same analysis for *S. cerevisiae* showed that amino acid sequences of this organism had about three elementary charges at the amino terminal end (data not known), indicating that the net charges of the main bodies of the proteins, except for the amino terminal regions, are neutral when half of the histidine residues are positively charged.

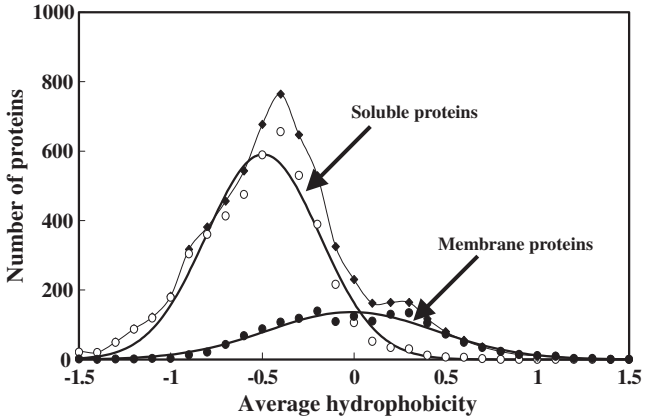

Figure 2. The distribution of average hydrophobicity of all proteins in the proteome of *S. cerevisiae*. The distribution is represented by two Gaussian distributions corresponding to the soluble and membrane proteins.

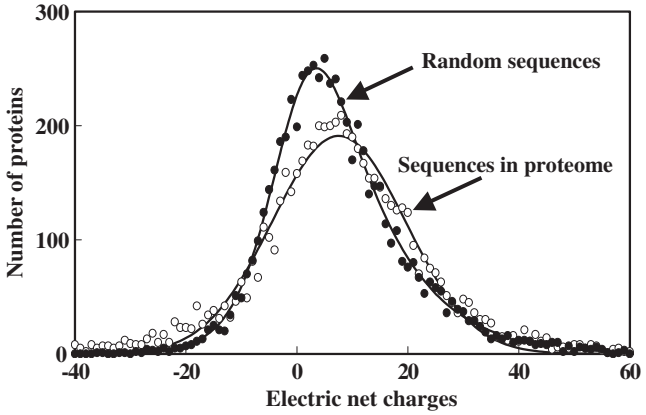

Figure 3. Comparison between the charge distribution for the *S. cerevisia* proteome and for random sequences. In the analysis of random sequences, the size distribution and composition of charged residues were kept the same as those for the *S. cerevisiae* proteome. The broadening of the distribution indicates a positive correlation of charges with sequences.

Despite a large variety of structures and functions, the entire set of proteins in a proteome appear to act together toward a common target of total net charges, otherwise the charge distribution would assume a more complicated shape. In fact, the distribution of the average hydrophobicity of all the proteins has two separate peaks, corresponding to the two classes of soluble and membrane proteins (figure 2) [7, 8]. Nucleotide sequences in a genome and amino acid sequences in a proteome are under mutation pressure during the evolutionary process. Despite this pressure, the average hydrophobicity of proteins exhibits a double peak. The functional importance of membrane proteins may explain why double peaks are maintained during the evolutionary process, indicating an analogous reason for the single Gaussian distribution of the net charges.

The distribution shift shown in figure 1 caused by the change in the dissociation condition of His indicates a reason for the single Gaussian distribution of net charges. Because the pK

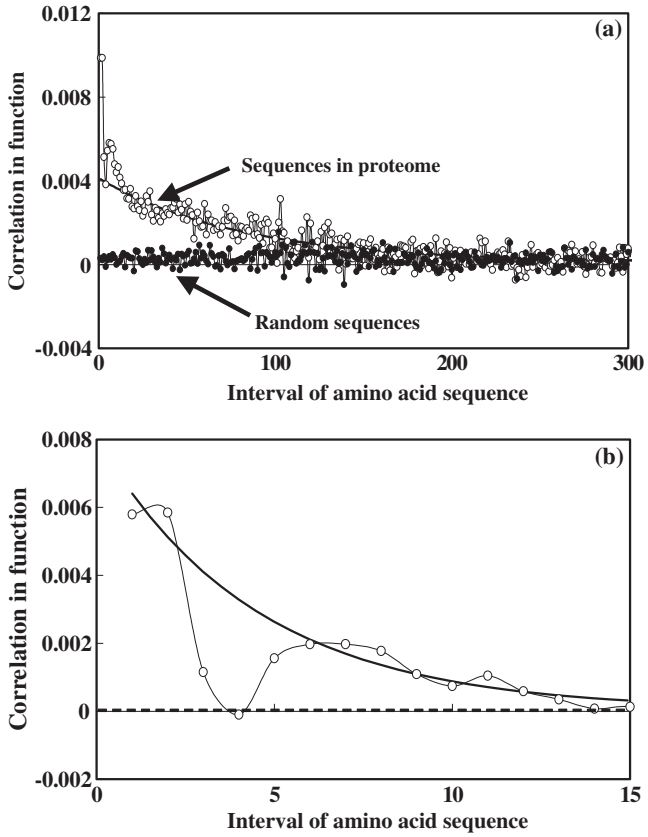

Figure 4. Autocorrelation function of all amino acid sequences in the *S. cerevisiae* proteome. The solid curve represents the exponential curves of the correlation function. The curve in the region above 20 residues showed a correlation length of 81 residues (a), whereas the corresponding length below 20 residues was 4.5 residues (b). A significant dip of the correlation function was observed at intervals of 3 and 4 residues.

of His is close to neutral (7.6), the degree of protonation depends upon the environment of the cell. When the degree of protonation changed from 0 to 1, the distribution mean varied from 0.2 to 7.6 elementary charges. Considering that the amino terminal end had a positive charge of about +3, many of which are cut off by signal peptidases, the entire set of proteins in a cell must exert buffering action around neutral pH.

A Gaussian distribution of net charges represents the randomness of the charges in amino acid sequences. However, it does not necessarily mean that the sequence of amino acids is completely random. Figure 3 shows the charge distribution for the *S. cerevisiae* proteome together with the distribution of random sequences generated by computer while maintaining the amino acid composition as well as the size distribution of sequences. The distribution for a proteome is much broader than that for completely random sequences. This difference is reasonable from a biological viewpoint. Each protein has its own function, which is determined by its 3D structure. Since the 3D structure is formed by a specific amino acid sequence, the sequence of charges should also be specific and not random. Therefore, the broad Gaussian distribution is a very important feature of a proteome, which should correlate with the sequences of charges in a proteome.

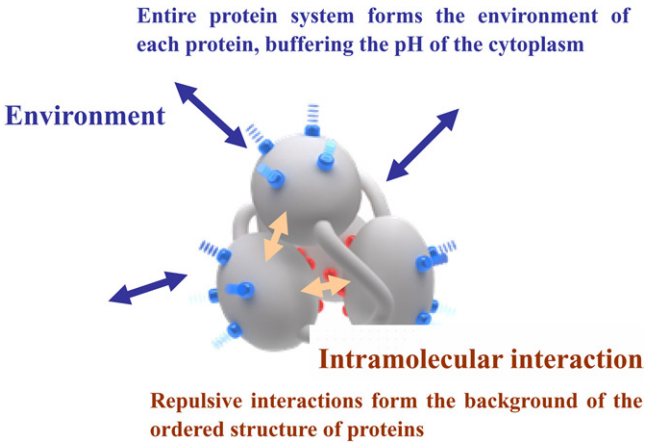

Figure 5. Two aspects of charge distribution of proteins at the proteome level: intramolecular interactions and interactions with the environment.

(This figure is in colour only in the electronic version)

We calculated the autocorrelation function of all amino acid sequences of *S. cerevisiae* proteome using equation (2). The autocorrelation function was also calculated for random sequences in which the amino acid composition and the size distribution of sequences were kept the same as the amino acid sequences in the *S. cerevisiae* proteome. Figure 4(a) shows the autocorrelation function of charges for the real proteome compared to that for the random sequences. The random sequences showed only a small constant term throughout the interval from 1 to 300 residues. The small positive correlation in the random sequences is due to the unbalanced charge totals of the amino acid composition: the ratio of positively charged residues (0.139) was slightly larger than that of negative ones (0.123). When we generated random sequences with the same amount of positive and negative charges, as a control simulation, the positive correlation completely disappeared. In contrast, the autocorrelation function of charge sequences of the *S. cerevisiae* proteome showed significant positive correlation. Two correlation lengths were observed: 81 residues for the interval region above 20 residues and 4.5 residues for the region below 20 residues. Another characteristic length was 3 or 4 residues at which a significant dip is observed, which suggests that helical structures in proteins are stabilized by the electric attraction at this interval (figure 4(b)).

4. Discussion

Analyses of charge distributions at the genome scale revealed two collective properties of the *S. cerevisiae* proteome. First, the distribution of net charges of all proteins was nearly neutral. Thus, the amino acid sequences in a proteome are designed so the whole system acts as a pH buffer. Second, the autocorrelation function of electric charges in all proteins showed a positive correlation as long as 81 residues, indicating that electric repulsion is dominant in this protein system. Both properties are significant on the scale of the proteome (figure 5).

If the electric charges in proteins function as a pH buffer, the correlation between the sequence of charges and the 3D structure or the function should be low. In fact, reports indicate that this type of correlation is unusual [9]. Therefore, the buffering action of proteins in a cell is reasonable from a physical viewpoint. However, a cell contains many other molecules, such as DNA, RNA, and lipids. Therefore, further experimental studies are required to reveal the contribution of proteins to the buffering action of a whole cell.

The role of repulsion in the structural formation process is difficult to understand fully, because most proteins conform to a globular shape with complicated ordered structure. The present work showed the importance of repulsion in the protein structures, but the globular shape of proteins has to be formed by the balance of attraction and repulsion, which needs to be studied for understanding the mechanism of protein folding.

Acknowledgment

This work was partly supported by Nagoya University 21st century COE 'Frontiers of Computational Science'.

References

[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P 2002 *Molecular Biology of the Cell* 4th edn (New York: Garland Science) chapter 5, pp 235–98
[2] Berg J M, Tymoczko J L and Stryer L 2002 *Biochemistry* 5th edn (San Francisco, CA: Freeman) chapter 3, pp 41–76
[3] Ke R and Mitaku S 2004 *Chem-Bio. Informatics J.* **4** 101
[4] Sear R P 2003 *J. Chem. Phys.* **118** 5157
[5] Goffeau A *et al* 1997 *Nature* **387** 5
[6] ftp://ftp.ncbi.nih.gov/genbank/genomes/S_cerevisiae/
[7] Mitaku S, Ono M, Hirokawa T, Boon-Chieng S and Sonoyama M 1999 *Biophys. Chem.* **82** 165
[8] Hirokawa T, Boon-Chieng S and Mitaku S 1998 *Bioinformatics Appl. Note* **14** 378
[9] Nakamura H 1996 *Q. Rev. Biophys.* **29** 1