Characterizing the Microenvironment Surrounding Phosphorylated Protein Sites

Shi-Cai Fan and Xue-Gong Zhang*

The Key Laboratory of the Ministry of Education for Bioinformatics, Department of Automation, Tsinghua University, Beijing 100084, China.

Protein phosphorylation plays an important role in various cellular processes. Due to its high complexity, the mechanism needs to be further studied. In the last few years, many methods have been contributed to this field, but almost all of them investigated the mechanism based on protein sequences around protein sites. In this study, we implement an exploration by characterizing the microenvironment surrounding phosphorylated protein sites with a modified shell model, and obtain some significant properties by the rank-sum test, such as the lack of some classes of residues, atoms, and secondary structures. Furthermore, we find that the depletion of some properties affects protein phosphorylation remarkably. Our results suggest that it is a meaningful direction to explore the mechanism of protein phosphorylation from microenvironment and we expect further findings along with the increasing size of phosphorylation and protein structure data.

Key words: protein phosphorylation, microenvironment, modified shell model, rank-sum test

Introduction

Protein phosphorylation is a ubiquitous post-translational modification occurring in either the cytosol or the nucleus of the cell, which is involved in many fundamental cellular processes, such as metabolism (*1*), apoptosis (*2*), cell signaling, and cellular proliferation (*3*). It is estimated that about 30%–50% of eukaryotic proteins undergo phosphorylation (*4*). Therefore, to investigate the mechanism of protein phosphorylation will be fairly useful to understand various protein functions and signal transduction pathways.

Biochemically, protein phosphorylation includes a transfer of a moiety of phosphate from adenosine triphosphate to the hydroxyl of acceptor residue, regulated by protein kinases (*5*). There are mainly three acceptor amino acids, namely serine (S), threonine (T), and tyrosine (Y), and many kinases could recognize substrates of both S and T sites (*6*).

Although the discovery of protein phosphorylation can be ascended to the fifties of 20$^{\text{th}}$ century, its mechanism still needs to be further studied due to its high complexity. In the early days, the investigation was carried out in experimental methods, which were accurate but hard and expensive. Then several computational methods were contributed to the field, including neural network (*5*), C4.5 (*7*), support vector machine (SVM; ref. *8*), *etc.*, all of which were proposed to explore protein phosphorylation based on sequences around phosphorylated sites. As we know, protein phosphorylation is a process that several molecules interact with each other in the space, and positional correlation in the sequence standpoint may not reflect the truth. For example, amino acids neighboring in the space may be distant in sequence interval. Consequently, the conclusions extracted from protein sequences may be not completely reliable.

In order to investigate the phosphorylation mechanism more directly, we propose to research from the microenvironment around phosphorylated and non-phosphorylated sites. As Altman's shell model (*9*, *10*) that accumulates the property distribution of each shell around a site was not applicable to our problem, we adopted a modified shell model that accumulated the spatial distribution of 80 biophysical and biochemical properties around a site at a distance range of 2–16 Å as a whole. As a result, we obtained some significant properties in the specified region by using the rank-sum test, such as the lack of some classes of residues, atoms, and secondary structures. Among all the properties, some are consistent

*** Corresponding author.**
E-mail: zhangxg@tsinghua.edu.cn

with the findings based on protein sequences, some are new, while others are somewhat different. We suspect that the depletion of some properties around sites may be more important than the enrichment. In a word, our method provides a new direction to investigate protein phosphorylation and we expect further findings when the size of phosphorylation and protein structure data becomes larger.

Results

After obtaining the structure data of positive and negative samples (sites and non-sites), we accumulated the spatial distribution of 80 biophysical and biochemical properties with respect to S, T, and Y sites by using our modified shell model, and adopted a standard nonparametric test of significance (the Mann-Whitney rank-sum test; ref. *11*) to compare the distribution. We listed in Table 1 the ten most significantly differential distributed properties for S, T, and Y sites, respectively, that is, the ten properties with the lowest p-value (p-value < 0.05), which were defined as our candidate properties.

In order to test whether these properties appeared randomly, we repeated 1,000 times of permutation on the samples followed by the rank-sum test, and counted the frequency of each candidate property (Table 1). From the result, we can see that most of the

Table 1 Significant Properties of Serine (S), Threonine (T), and Tyrosine (Y) Sites

Site	Property	p-value	Frequency in randomicity test	Frequency in sensitivity test	Significant status
Serine (S)	Residue-name-is-Ile*	0.00058	34	999	low
	Ring-system*	0.00102	37	931	low
	Mobility*	0.00118	24	997	low
	Residue-name-is-Phe*	0.00129	41	944	low
	Atom-name-is-C*	0.00156	28	996	low
	Residue-class2-is-basic*	0.00239	46	921	low
	Atom-type-is-CT	0.00399	35	716	low
	Atom-name-is-N	0.00412	52	806	low
	Vdw-volume	0.00417	34	777	low
	Partial-charge	0.00582	50	328	high
Threonine (T)	Residue-class1-is-hydrophobic*	0.00012	31	1,000	low
	Residue-name-is-Val*	0.00013	38	994	low
	Residue-class2-is-nonpolar*	0.00026	28	973	low
	Atom-name-is-C*	0.00037	26	1,000	low
	Atom-type-is-CT*	0.00040	33	998	low
	VDW-volume	0.00060	54	920	low
	Atom-type-is-N	0.00060	29	779	low
	Amide	0.00070	55	819	low
	Atom-name-is-any	0.00070	22	598	low
	Atom-type-is-O	0.00070	30	410	low
Tyrosine (Y)	Secondary-structure2-is-beta*	0.012	41	1,000	low
	Charge*	0.013	34	990	high
	Residue-name-is-Cys*	0.017	44	977	low
	Residue-name-is-Pro*	0.017	46	1,000	low
	Atom-type-is-N*	0.019	37	1,000	low
	Residue-class1-is-hydrophobic*	0.023	43	967	low
	Residue-name-is-Asp	0.023	47	580	high
	Residue-name-is-Leu	0.025	52	828	low
	Secondary-structure1-is-strand	0.029	41	689	low
	Atom-type-is-O	0.030	47	310	high

*Significant properties.