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.

candidate properties appeared less than 50 times, and such properties are unlikely to be random.

We continued to test whether these properties were sensitive to the negative sample size, for the number of non-sites was far more than that of the sites. By clustering the negative samples to 4–5 times as many as positive samples for 1,000 times, we obtained 1,000 groups of samples; each was composed of a positive sample and one of the 1,000 negative samples. Then we implemented the rank-sum test on the 1,000 groups of data, respectively, and counted the frequency of each candidate property (Table 1). From the result, we can see that some properties are very stable, and we will ignore the unstable properties in the end.

Considering both the randomicity and sensitivity tests, we selected the properties that appeared less than 50 times in the randomicity test and more than 900 times in the sensitivity test to characterize the microenvironment of S, T, and Y sites, and marked them in Table 1. We can see that there are 6, 5, and 6 significant properties to S, T, and Y sites, respectively.

Discussion

Based on these significant properties, we have the following conclusions: (1) Around S sites, it is deficient in residues Ile and Phe, which are characterized as nonpolar and hydrophobic that are consistent with another two properties, ring-system and residue-class2-is-basic. In addition, mobility and atom-name-is-C are also deficient around S sites. (2) Around T sites, the lack of residue Val is correlated with the depletion of hydrophobic and nonpolar; meanwhile, atom-type-is-CT and atom-name-is-C are underrepresented, and VDW-volume is also much smaller. From such results, we can see that there are some similar features between S and T sites, identical with the report that many protein kinases can recognize substrates of both S and T sites (*6*). (3) Around Y sites, there is only one enriched property, namely charge, among all the significant properties. Meanwhile, it is notably lack of two residues, Cys and Pro, which are both neutral. Furthermore, hydrophobic, the secondary structure of beta, and atom-type-is-N are significantly absent around Y sites.

Among all the properties we investigated, some are the same with the findings based on the protein sequences around sites, such as the depletion of Ile, Phe, and Val that are nonpolor and hydrophobic, the depletion of Cys and Pro that are neutral, and the enrichment of charge (*12*); some are new findings, such as the lack of atom-name-is-C/N, atom-type-is-CT/N, and the smaller VDW-volume; others are somewhat different, and the most difference is that our conclusion tends to highlight the depletion of some properties rather than the enrichment emphasized in methods based on protein sequences, for nearly all of the properties we extracted are significantly low by using the rank-sum test. Therefore, we propose that the depletion of some properties around sites has much more effect on protein phosphorylation.

In general, these biophysical and biochemical properties provide an insight that may be functionally related with the mechanism of protein phosphorylation, and such analysis may be important for protein engineering application.

In addition, after obtaining these significant properties, we implemented site prediction using SVMs (*8*), and the prediction accuracy of S, T, and Y sites were all more than 80% by cross-validation. Although our accuracy is not higher than that of sequence-based methods (*13*), our sample size is much smaller. We expect that it will be significative to explore the mechanism of protein phosphorylation in the view of microenvironment along with the increasing size of phosphorylation and protein structure data.

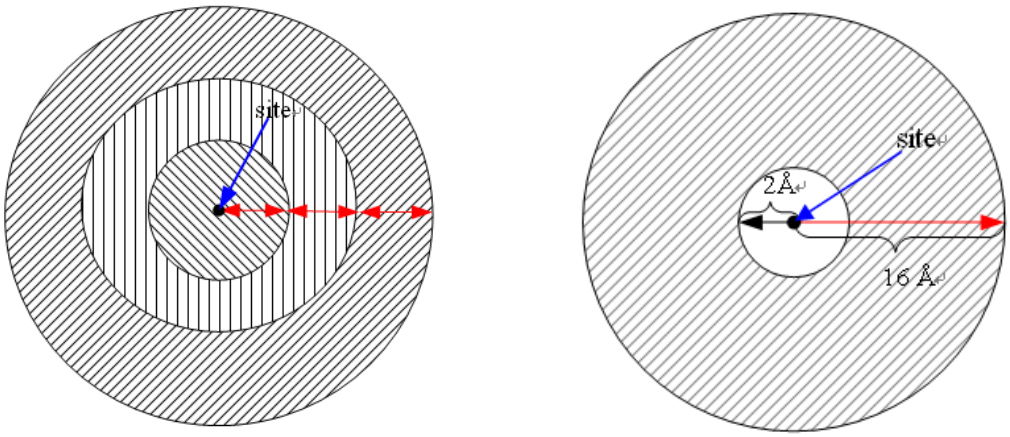
Materials and Methods

Datasets

We obtained the dataset of phosphorylation sites from Phospho.ELM (version 3.0; ref. *14*) containing protein sequences, phosphorylated sites (S, T, and Y), and corresponding protein kinases. All of them have been validated by experiment. As there were no exact negative samples, we extracted S, T, and Y of the same proteins reported in Phospho.ELM without annotating them as phosphorylated sites to be our negative samples. In addition, we obtained the dataset of protein structures from the Protein Data Bank (PDB; ref. *15*). For the protein that has more than one record in PDB, we selected the one with the highest resolution. Only the sites reported in both databases were extracted as our samples. The sample sizes of S, T, and Y are shown in Table 2.

Table 2 Sample Data of S, T, and Y Reported in Both Phospho.ELM and PDB Databases

Amino acid residue	Positive sample size	Negative sample size
Serine (S)	42	433
Threonine (T)	19	232
Tyrosine (Y)	39	203



A **B**

Fig. 1 A. Altman's shell model that accumulates the property distribution of each shell around a site. **B**. The modified model that accumulates the property distribution around a site at a distance range of 2–16 Å as a whole.

Properties

We used 80 biophysical and biochemical properties to characterize the microenvironment around S, T, and Y sites as listed below:

1. Atom-based properties: atom type, hydrophobicity, charge, and charge-with-His.

2. Chemical group-based properties: hydroxyl, amide, amine, carbonyl, ring-system, and peptide.

3. Residue-based properties: residue type, hydrophobicity classifications 1 and 2.

4. Secondary structure-based properties: secondary structure classifications 1 and 2.

5. Other properties: VDW-volume, B-factor, mobility, and solvent accessibility.

These properties include almost all the properties within a biomolecular structure and have been used to characterize the microenvironment of other sites (*9*, *16*).

Model

Altman and his colleagues (*9*, *10*) adopted a shell model to characterize the microenvironment surrounding calcium binding sites, whose main idea was to accumulate the property distribution of each shell around a site (Figure 1A). In this study, we applied a modified shell model on our data, which accumulated the property distribution around a site at a distance range of 2–16 Å as a whole (Figure 1B). Our reason is as follows: when we applied the original shell model on our data, the significant properties in each shell extracted by the rank-sum test were sensitive to the negative sample size, and we ascribed it to that the sites were not affected by the properties in each shell separately but the overall properties of all the shells. Therefore, we modified the model and extended the distance from 2 Å to 16 Å for most significant properties between sites and non-sites concentrated in such regions.

Randomicity test

In order to test whether these significant properties appeared randomly, we implemented 1,000 times of permutation on the positive and negative samples. The permutation process was as follows: on the assumption that there were N_p positive and N_n negative samples, we first mixed the positive and negative samples together ($N_p + N_n$ samples in all), then drew N_p positive and N_n negative samples randomly. After repeating the permutation for 1,000 times, we obtained 1,000 groups of data.

Sensitivity test

In order to test whether these significant properties were stable, we clustered the number of negative samples to 4–5 times as many as that of positive samples using the BLOSUM62 matrix (*17*) by setting proper similarity cutoff, and selected one sample from each group to form negative samples. As there is inevitably a certain randomicity when selecting the first sequence to begin clustering, the result of each clustering may be different to some extent, so we repeated the process for 1,000 times and obtained 1,000 groups of negative samples.

Acknowledgements

This work was partly supported by the National Key Technologies R&D Program (No. 2004BA711A21) and the National Natural Science Foundation of China (No. 60275007 and 60234020).

References

1. Salway, J.G. 1999. *Metabolism at a Glance* (second editon). Blackwell Publishing Ltd., Oxford, UK.
2. Rang, H.P., *et al.* 1999. *Pharmacology* (fourth edition). Churchill Livingstone, Edinburgh, UK.
3. Matthews, H.R. 1995. Protein kinases and phosphatases that act on histidine, lysine, or arginine residues in eukaryotic proteins: a possible regulator of the mitogen-activated protein kinase cascade. *Pharmacol. Ther.* 67: 323-350.
4. Hardie, D.G. (ed.) 1999. *Protein Phosphorylation: A Practical Approach.* Oxford University Press, New York, USA.
5. Blom, N., *et al.* 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* 294: 1351-1362.
6. Pinna, L.A. and Ruzzene, M. 1996. How do protein kinases recognize their substrates? *Biochim. Biophys. Acta* 1314: 191-225.
7. Berry, E.A., *et al.* 2004. Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. *Comput. Biol. Chem.* 28: 75-85.
8. Kim, J.H., *et al.* 2004. Prediction of phosphorylation sites using SVMs. *Bioinformatics* 20: 3179-3184.
9. Bagley, S.C. and Altman, R.B. 1995. Characterizing the microenvironment surrounding protein sites. *Protein Sci.* 4: 622-635.
10. Wei, L. and Altman, R.B. 1998. Recognizing protein binding sites using statistical descriptions of their 3D environments. *Pac. Symp. Biocomput.* pp.497-508.
11. Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics* 1: 80-83.
12. Iakoucheva, L.M., *et al.* 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32: 1037-1049.
13. Vapnik, V.N. 1995. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, USA.
14. Diella, F., *et al.* 2004. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 5: 79.
15. Berman, H.M., *et al.* 2000. The protein data bank. *Nucleic Acids Res.* 28: 235-242.
16. Liang, M.P., *et al.* 2003. WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res.* 31: 3324-3327.
17. Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915-10919.