Max-Planck-Institut
für Molekulare Pflanzenphysiologie

GabiPD – Access to natural variation in Arabidopsis and effects on phosphorylation patterns

gabi
Genomanalyse
im biologischen
System Pflanze

Sabrina Kleessen[1], Diego Mauricio Riaño-Pachón[1], Jost Neigenfind[1], Pawel Durek[4], Dirk Walther[1], Joachim Selbig[3], Waltraud Schulze[2] & Birgit Kersten[1]

Contact: gabipd@mpimp-golm.mpg.de

[1] Bioinformatics Group, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany
[2] Signaling Proteomics Group, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany
[3] Institute for Biochemistry and Biology, University of Potsdam, c/o MPIMP, Am Mühlenberg 1, 14476 Potsdam, Germany
[4] Institute of Pathology, Universitätsmedizin Charite, Chariteplatz 1, 10117 Berlin, Germany

Introduction

Protein phosphorylation is an important post-translational modification which influences virtually every aspect of dynamic cellular behavior by controlling, e.g. cellular signaling. Site-specific phosphorylation of amino acid residues serine, threonine and tyrosine can have profound effects on protein structure, activity, stability and the interaction with other biomolecules. Hence, the systematic analysis of protein phosphorylation by phosphoproteomic and bioinformatic approaches is of great importance in understanding cellular functions [1]. In *Arabidopsis thaliana*, recent progress in the identification of phosphosites (psites) has prompted the creation of dedicated web-resources, e.g. PhosPhAt, and has made it possible to develop Arabidopsis-specific predictors of phosphosites [2]. Moreover, recent advances in polymorphism typing in *Arabidopsis thaliana* paved the way to study the effects of these polymorphisms, especially of SNPs, on other genome-wide features, such as phosphorylation patterns.

Number of SNPs

Recent developments in genomics have allowed the analysis of polymorphic sites in a genome-wide scale in *A. thaliana*. SNPs have been identified through the comparison of 100 *A. thaliana* accessions at the genomic level in different studies (Table 1, Figure 1).

Table 1: Number of SNPs detected in each of the *A. thaliana* studies. Results of SNP mapping onto cDNAs/CDS and number of synonymous/non-synonymous substitutions caused by SNPs.

Dataset	Number of non-redundant SNPs in this study	Number of non-redundant SNPs mapping onto cDNAs	Number of non-redundant SNPs mapping onto CDS	Number of non-redundant SNPs causing at least one non-synonymous substitution	Number of non-redundant SNPs always causing synonymous substitutions
Nordborg [4]	20667	9251	8023	4047	3975
Clark [5]	637522	263718	227497	109709	117788
Ossowski [6]	860154	220984	174559	84400	90159
TOTAL	1247284	382770	315039	156034	159004

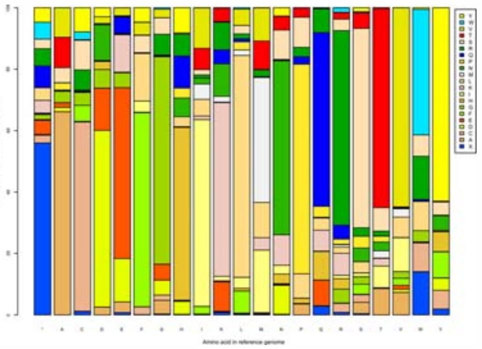

Figure 1: Pattern of amino acid substitutions in polymorphic sites. For every position each amino acid substitution was counted only once.

Phosphosites and hotspots

Experimental psites of different phosphoproteomic studies in *A. thaliana* were taken from PhosPhAt (version 3.0) [2] yielding a total of 7178 unambiguously identified phosphopeptides identified in 4252 protein-coding loci. The phosphorylation pattern displayed a distribution as presented in Figure 2. In addition, we used proteome-wide predicted *A. thaliana* psites to get a more global view. This dataset, also taken from PhosPhAt [2], comprises 75296 high confidence predicted phospho-peptides (score≥1), identified in 21711 protein-coding loci.

Based on experimental and predicted psites, we computed potential hotspots of phosphorylation, where a hotspot was defined as a sequence window of a given length containing a significantly increased number of phosphorylated S, T, and Y sites (Figure 2). For psites as well as for hotspots, our analyses confirmed a previously suggested trend of psites to be located outside conserved protein domains.

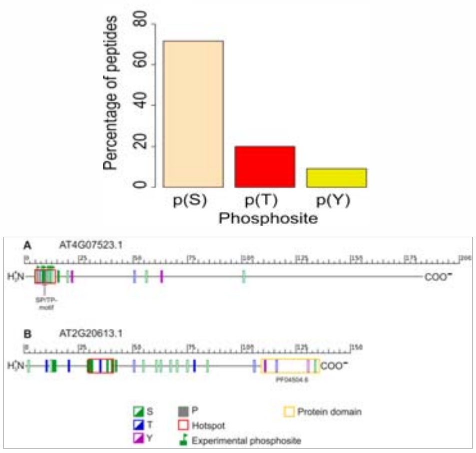

Figure 2: Percentage of peptides in experimental phosphosite data (on top). Hotspots of phosphorylation computed in two *A. thaliana* proteins based on experimental (on bottom, A) or predicted phosphosites (on bottom, B). Solid colored boxes represent predicted phosphosites (score>0).

Distribution of psites

We were then interested in the distribution of the psites across the different individual *A. thaliana* proteins. For both datasets, experimental and predicted psites, most of the proteins contained only few psites whereas a few proteins were phosphorylation hubs, i.e., included a large number of psites. This kind of distribution is compatible with a power-law, which had been shown to apply for the distribution of the number of psites per protein in other organisms [3]. The power-law-like distribution still holds after the total number of S, T, and Y residues is used to normalize the cumulative probabilities as shown in Figure 3.

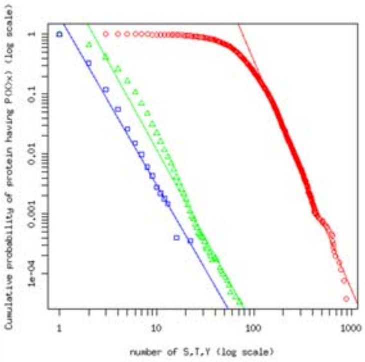

Figure 3: Cumulative distribution of the frequency of the number of phosphosites per protein, the probabilities have been normalized to take into account the total number of S, T, Y residues in proteins of each class, i.e., with 1, 2, 3 etc. phosphosites. The series with blue squares is the distribution in experimentally identified phosphosites, the green triangles is the distribution in predicted phosphosites. This type of distribution suggests a rich-get-richer process for the accumulation of phosphosites per protein [3]. We observed a similar distribution for the total number of S, T or Y residues per proteins (red circles).

Loss and gain of psites

Based on the current dataset of experimental psites we identified 86 experimental psites in 86 gene models that are lost by a nsSNP (non-synonymous SNP), i.e., where the amino acid at the psite is exchanged by any other amino acid.

Using the dataset of predicted high confidence phosphorylation sites for a more global analysis, we found that 1114 proteins with predicted phosphosites could potentially lose a phosphosite in at least one of the Arabidopsis accessions studied.

To identify over- or underrepresented biological functions among the proteins containing a loss or gain of predicted phosphorylation site, these proteins were tested against a reference set that contains all proteins with at least one predicted phosphosite (Figure 4). Over- and under-representation analysis of GOSlim terms was carried out using BiNGO (http://www.psb.ugent.be/cbd/papers/BiNGO/).

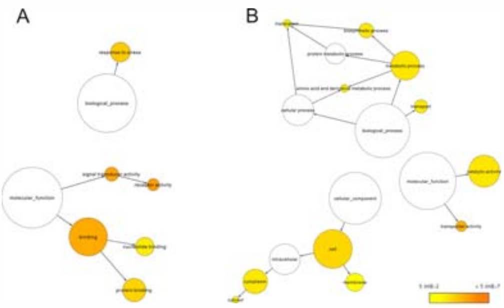

Figure 4: Overrepresented (A) and underrepresented (B) GO terms in AGIs with predicted gain or loss phosphosites (by nsSNP). The dataset was compared with a reference set that comprised all proteins containing at least one high confidentially predicted phosphosite (score≥1).

Integration into GabiPD

The integrated SNP data is publicly available via the GABI Primary Database (GabiPD; http://www.gabipd.org/) [7].
More than 380000 non-redundant SNPs are currently available through the *A. thaliana* Gene GreenCards
(Figure 5).

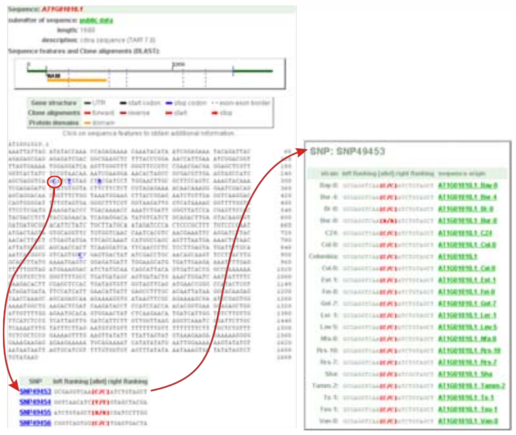

Figure 5: SNPs integrated into an Arabidopsis Gene GreenCard in GabiPD

Conclusions

The phosphoproteomic dataset compiled in this work represents the currently largest phosphosite dataset in Arabidopsis.
By mapping nsSNPs onto phosphosites we identified losses and gains of phosphosites, which can be important in different adaptive responses of the natural accessions in their different environments. Especially proteins involved in signaling and stress response were affected by changes (gain or loss) of predicted phosphosites, whereas proteins involved in metabolism, catalytic activity and biosynthesis were less affected.

Funding

This study was supported by a grant from the German Federal Ministry of Education and Research (BMBF grant 0315046) and by the Max-Planck-Society.

Bundesministerium
für Bildung
und Forschung

References

[1] Kersten B, Agrawal GK, Durek P, Neigenfind J, Schulze W, Walther D, Rakwal R (2009) Plant phosphoproteomics: An update. Proteomics 9(4):964-88.

[2] Durek P, Schmidt R, Heazlewood J, Jones A, MacLean D, Nagel A, Kersten B, Schulze W (2010) PhosPhAt: The Arabidopsis thaliana phosphorylation site database. An update. Nucleic Acids Research 38 (Database issue): D828-34.

[3] Yachie N, Saito R, Sugahara J, Tomita M, Ishihama Y (2009) In silico analysis of phosphoproteome data suggests a rich-get-richer process of phosphosite accumulation over evolution. Mol Cell Proteomics 8(5):1061-1071.

[4] Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R et al (2005) The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3(7):e196.

[5] Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA et al (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 317(5836):338-342.

[6] Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 18(12):2024-2033.

[7] Riano-Pachon DM, Nagel A, Neigenfind J, Wagner R, Basekow R, Weber E, Mueller-Roeber B, Diehl S, Kersten B (2009) GabiPD: the GABI primary database--a plant integrative 'omics' database. Nucleic Acids Research 37(Database issue): D954-959.