Journal of
Proteomics & Bioinformatics

Edwards et al., J Proteomics Bioinform 2012, 5:4
DOI: 10.4172/jpb.1000220

Rapid Communication Open Access

ReportSites - A Computational Method to Extract Positional and Physico-Chemical Information from Large-Scale Proteomic Post-Translational Modification Datasets

Alistair VG Edwards[1,2]*, Gregory J Edwards[4], Martin R Larsen[2] and Stuart J Cordwell[1,3]

[1]Discipline of Pathology, School of Medical Sciences, The University of Sydney, Australia
[2]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, Odense, Denmark
[3]School of Molecular Bioscience, The University of Sydney, Australia
[4]Port Jackson Bioinformatics, Sydney, Australia

Abstract

Background: Extracting biological meaning from proteomic datasets containing post-translational modification is a central challenge of large scale proteomics and systems biology. We report the generation of a new program (ReportSites) to precisely identify the location and local chemical environment of a particular amino acid residue, or group of residues, within large-scale proteomic data sets, using peptide sequences characterized by mass spectrometry combined with protein sequence databases. The program is ideally suited to distilling regional and spatial information from post-translational modification data sets, wherein patterns of sequence surrounding processed sites may reveal more about the functional and structural requirements of the modification and the biochemical processes that regulate them.

Results: We developed ReportSites using a test set of phosphoproteomic data from rat myocardium that contains approximately eleven thousand unique phosphorylation sites. These sites were used to identify patterns associated with site location (spatial sequence information) within the context of the complete protein sequence, and two selected aspects of the immediate physico-chemical environment (local p*I* and hydrophobicity). These were then also compared to corresponding values extracted from the full database to allow comparison of phosphorylation trends.

Conclusions: ReportSites enabled physico-chemical aspects of protein phosphorylation to be deciphered in a test set of eleven thousand phospho sites. Basic properties of modified proteins, such as site location in the context of the complete protein, were also documented. This program can be easily adapted to any post-translational modification (or, indeed, to any defined amino acid sequence), or expanded to include more descriptive factors (such as modification of binding domains or protein structure).This makes it a versatile tool with the potential to aid in revealing new aspects of post-translational modification distribution. The code is freely available from the authors upon request and is accessible online.

Keywords: Protein modification; Software; Phosphorylation; Distribution

Abbreviations: KEGG: Kyoto Encyclopaedia of Genes and Genomes; PTM: Post-Translational Modification; STRING: Search Tool for Recurring Instances of Neighboring Genes

Introduction

Much effort is currently expended in the search for biological meaning in large-scale proteomic datasets. The sheer volume of data, even in the context of quantitative changes associated with a specific experimental question, means it is not trivial to transition from a list of protein identifications to hypothesis generation regarding biological impact. The problem is substantially compounded by the generation of higher order information regarding post-translational modification (PTM) of proteins, where the effect on the predicted function of the identified protein must be considered in the context of the precise location of the modification. Many investigators choose to use databases and web-based applications such as KEGG (www.genome.jp/kegg/) and STRING (string-db.org) to extract interpretations from modificomic data, particularly for signalling studies, by mapping PTM data on to known signalling pathways and interaction networks [1,2]. Such an analysis is heavily influenced by the accuracy of the relevant database and is also biased to the protein identification inferring a function for the identified PTM. A complementary approach is to apply a probabilistic method to describe ways in which the dataset over- or under-represents certain known characteristics (e.g. protein kinase motifs [3]). In a similar manner, it is likely to be informative to describe basic trends in PTM patterns associated with the local amino acid sequence environment of identified modification sites that may aid in determining their role in cellular behaviour under different biological conditions.

We have written a program in Perl that is able to eliminate redundancy in site reporting, document the precise site of modification in the context of the whole protein and record the physico-chemical environment of the modification site. For example, one can detect N- to C-terminal distributions of modification and simultaneously describe the local features of the sites detected (including p*I*, hydrophobicity

***Corresponding author:** Alistair VG Edwards, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, Odense, Denmark, E-mail: aled@bmb.sdu.dk

Received February 13, 2012; **Accepted** March 05, 2012; **Published** April 30, 2012

Citation: Edwards AVG, Edwards GJ, Larsen MR, Cordwell SJ (2012) ReportSites - A Computational Method to Extract Positional and Physico-Chemical Information from Large-Scale Proteomic Post-Translational Modification Datasets. J Proteomics Bioinform 5: 104-107. doi:10.4172/jpb.1000220

and motif/sequon) such that any positional trends or trends in any of the interrogated variables can be detected. This tool was then applied to a large-scale dataset (a phosphoproteomic study of rat myocardium) in both a site-oriented and protein-oriented manner. This does not attempt to replace probabilistic methods, rather providing a complementary tool to detect and describe trends in the distribution of various characteristics which may be relevant to a range of PTMs and which may be difficult to capture statistically due to low stoichiometry of modified species.

Materials and Methods

ReportSites is a command-line oriented Perl program of ~9000. Its central function is to overlay and merge the peptide sequences from MS/MS data into a nominated protein database and thereby eliminate redundant modified sites which are discovered more than once in peptide sequences from the same protein. In the dataset used these are phosphorylation sites, i.e. phospho Ser, Thr or Tyr. In the case of test data, the database used to develop the code was ipi.RAT.v3.52.fasta, which can be obtained from ftp://ftp.ebi.ac.uk/pub/databases/IPI/old/RAT. In order to develop and test this code, a large scale phosphoproteomic dataset was produced. Rat myocardial protein samples (produced under guidelines set out by The University of Sydney Animal Ethics Committee, approval number K20/6–2009/3/5078) were enriched for phosphopeptides using titanium dioxide essentially following the method of Larsen et al. [4] and analysed on an LTQ-OrbitrapXL (Thermo Scientific, San Jose, CA) using collision induced dissociation and electron transfer dissociation. Raw files were processed and searched using Proteome Discoverer (version 1.0; Thermo Scientific) and Mascot (version 1.12; www.matrixscience.com/) against an International Protein Index database (IPIRat, version 3.52). Peptide hits with a Mascot score lower than 20 were discarded. The code performs a number of data validity checks, and calculates statistics and distributions of phosphorylation sites, as well as p*I* and hydrophobicity around the sites. These are reported in a summary form and graphed in a Perl graphing module, and written to extensive CSV files for further analysis or graphing. The code has a linear flow from top to bottom, and its functions can be illustrated by a list of main actions (Supplementary Table 1). These functions include basic versions of existing analytical tools (e.g. motif/sequon oriented analysis) which can be used in conjunction with other aspects of the program if so desired. Sites of modification were denoted with a lower case character.

Results

Site analyses

Data produced from mass spectrometric analysis of phosphopeptide-enriched samples were reduced to a non-redundant set of phosphorylation sites using ReportSites, which produced a list of 10,882 unique phosphorylation sites. The distribution of these sites along the length of the identified proteins were analysed to observe any trends in the position of phosphorylation. In the test dataset, *N*- and *C*-terminal regions contained a higher proportion of phosphorylation sites compared to the internal regions of proteins (Figure 1), which is in line with some previous work showing a bias towards termini by kinases (e.g. in intermediate filament proteins [5]).

We then employed ReportSites to assess the local amino acid sequence hydrophobicity and p*I* surrounding the 10,882 phosphorylation sites. We defined 'local' as 6 amino acid residues both up- and downstream of the identified modification site, although this value can be altered to contain as large or small a region as is desired in a given processing run. Hydrophobicity was assessed on the scale according to Kyte and Doolittle [6], andregional pI was calculated using residue side chain values. ReportSites produced output charts of average hydrophobicity surrounding modification sites (Figure 2A), as well as charts of modification site hydrophobicity variation along

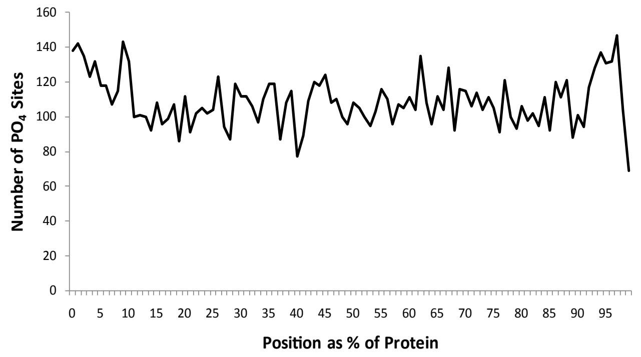

Figure 1: Positional distribution of observed phosphorylation sites in test data set (>10,000 observed phosphorylation sites). Position of detected phosphorylation sites displayed as a percentage of relevant protein length (*N*- to *C*-terminal) and showing preference for localization to the termini.

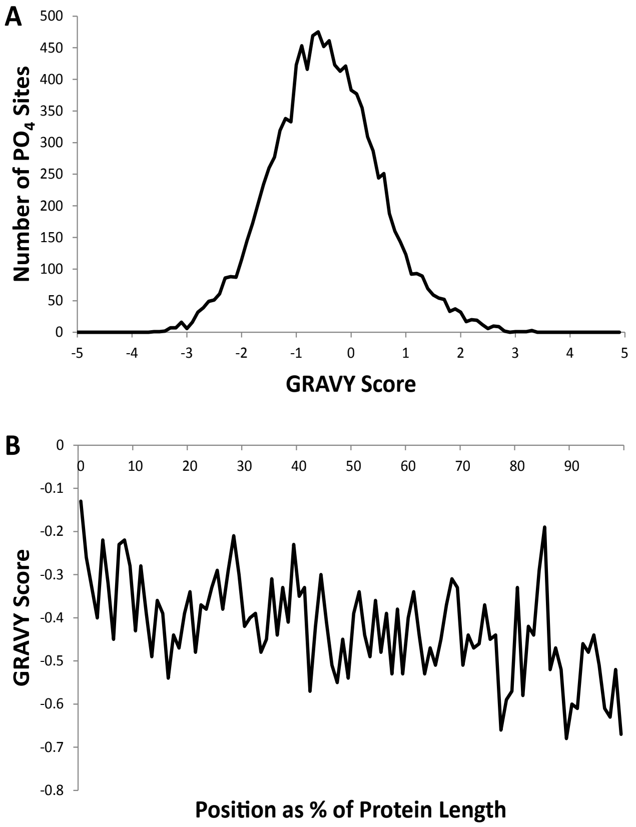

Figure 2: Hydrophobicity plots. (A) Average hydrophobicity of local environment (13 amino acid residues comprised of -6 to +6 around the phosphorylation site itself) surrounding phosphorylation sites. (B) Sliding window hydrophobicity plot across proteins and sites identified in dataset (expressed with reference to the position of the modification site position as a percentage of protein length, similar to Figure 1).

the length of all proteins identified using a sliding 13 residue window (Figure 2B). Similar output was generated for p*I* (Figure 3). It was also possible to perform these analyses for all instances of a particular amino acid residue, whether modified or not, in the given database (e.g. UniProt, IPI), such that the deviation of the experimental dataset from the sequence database may be clearly detected. We used ReportSites to perform such a database comparison for hydrophobicity, which in this case suggested that there was little difference in local hydrophobicity between identified phosphorylation sites and all sites in the IPI Rat database that could potentially be phosphorylated (Figure 4). Finally, since in eukaryotes phosphorylation occurs on serine, threonine and tyrosine residues, ReportSites can be used to examine the differences between identified phosphorylation sites on these residues with regard to any of the above-mentioned factors (e.g. hydrophobicity; Figure 5).

Discussion

We here report a simple software tool, ReportSites that can facilitate site-specific analysis of proteomic data to look for patterns of local sequence information based on physico-chemical properties. We envisage that such data will predominantly consist of large PTM-based mass spectrometry-based analyses, although the software is equally applicable to datasets containing information about ligand binding sites, structurally important residues, or alternative site-specific biological phenomena. We have chosen to limit the analysis to phosphorylation site p*I*, hydrophobicity, and site position, all factors of substantial biological importance [7] which may either determine the distributions of modifications or report new aspects of these distributions. ReportSites also allowed us to easily reduce a complex and overlapping list of modified peptide data to a non-redundant site catalogue, thus reducing bias caused by repetitive identification of the same site in replicate peptide sets (e.g. due to missed proteolytic cleavages, etc.). Based on analysis of our own data and that of others, it is our observation that non-redundant site counts are often up to 30% lower than estimates made in other ways. The chemical and physical environment surrounding a modification site can be examined with reference to the sites as a group or to the position within the identified proteins. In the current study, we used ReportSites to examine a large-scale (>10,000 sites) phospho proteomic dataset as an example and therefore defined Ser, Thr and Tyr as our amino acid residues of interest, but one could, for example, just as easily nominate asparagines in order to analyse *N*-linked glycosylation or deamidation, or other amino acids involved in different PTMs. Equally, the size of the analysis window (here 6 residues before and after the site of modification) can be altered to include as large or small region as required.

Our major aim was to produce a software tool and we therefore only superficially examined the results from our test dataset in a biological context. Furthermore, there are some shortcomings in the

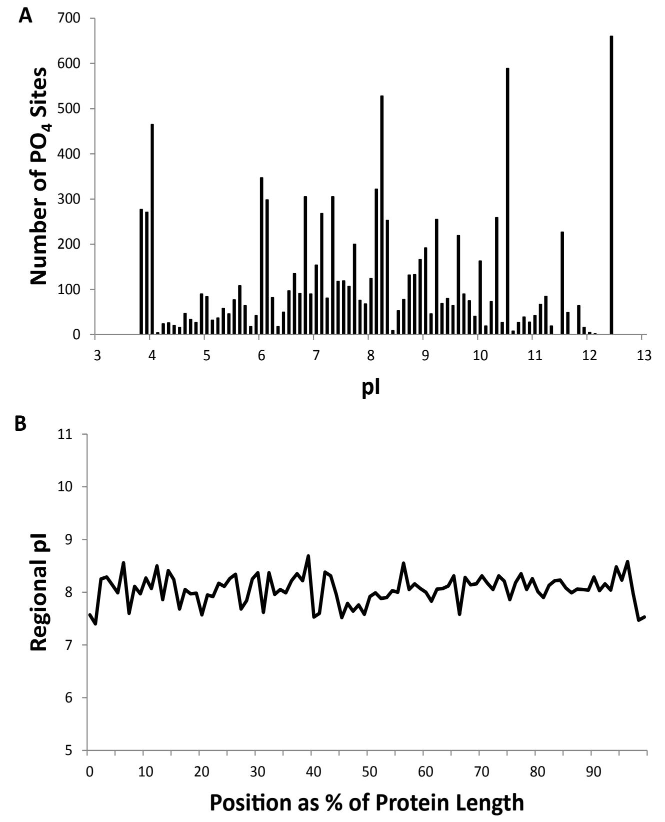

Figure 3: pI plots. (A) Average p*I* of local environment (13 amino acid residues comprised of -6 to +6 around the phosphorylation site itself) surrounding phosphorylation sites. (B) Sliding window p*I* plot across proteins and sites identified in dataset (expressed with reference to the position of the modification site position as a percentage of protein length, similar to Figure 1).

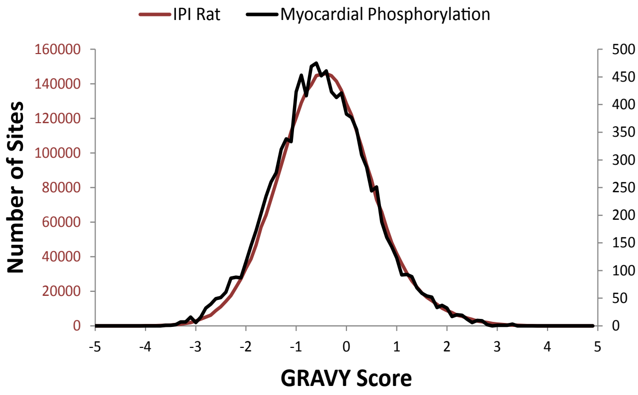

Figure 4: Hydrophobicity plots. Comparison of hydrophobicity distribution between sites identified in test dataset (myocardial phosphorylation; solid line) versus all possible sites in IPI Rat database (IPI Rat; based on serine, threonine and tyrosine sites; dashed line).

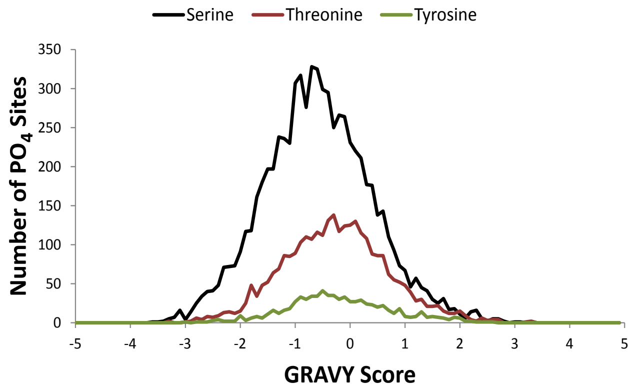

Figure 5: Residue hydrophobicity plots. Example plot of one variable (hydrophobicity) assessed separately for each type of residue (serine, threonine and tyrosine) in dataset (expressed over 13 residue windows).

data used to develop ReportSites (e.g. a lack of site scoring, in the manner of Ascore [8]). Several other tools are available to perform various analyses of proteomic data (e.g. STRING [9], NetworKIN [10]) that provide pathway analysis in a functional context. Fewer tools allow the user to visualise various physico-chemical aspects of protein and peptide structure (e.g. GPMAW[11], as well as several programs available through ExPASy [www.expasy.ch]), including calculations of protein p*I* / mass, hydrophobicity, functionally-associated motifs and trans-membrane predictions. Despite this, we are unaware of any freely accessible programs that allow interrogation of local peptide features with reference to a specific, user-defined central point while simultaneously presenting this information in the context of the broader experimental dataset and full proteome, if required. A good comparison is the motif-distilling software of Schwartz et al. [3], MotifX, which allows the user to set a target residue and assess the common patterns of residues, or motifs, surrounding them in a given dataset. ReportSites differs from MotifX in that it is not probabilistic and therefore the output is not affected by the background dataset defined by the user, and also in the wider range of variables that can be queried. The first of these points is particularly relevant as more modifications are being analysed at sub-stoichiometric levels which will complicate probabilistic identification.

The option to display all interrogated factors with reference to the portion of the relevant proteins in which they are found and also with reference to an entire database of similar information allows for two things: firstly, to easily discern positional trends of the modification or feature of interest (of interest in, for example, terminal phosphorylation-driven regulation of protein-protein interactions or protein translocation (e.g. [12]), and, secondly, to assess the deviation of the data from any ‘normal’ values – any substantial differences in modification patterns on a proteome-wide basis will in this way be detected. With the inclusion of more modifications for analysis (e.g. glycosylation and acetylation), the investigation of the interaction between these modifications also becomes possible, such that the distribution of a particular modification could be mapped relative to a second modification and the ‘cross-talk’ between them assessed.

Conclusions

Development of ReportSites allowed us to describe more clearly the location and physico-chemical trends of protein phosphorylation in rat myocardium. The variables chosen within ReportSites can be altered by the user (parameters to investigate, length of local environment, database, etc.). On a large enough proteomic dataset (now commonly produced), this may aid in revealing substantial and novel aspects of cellular control of post-translational modification, ligand binding and other functional properties. Therefore, we conclude that interrogation of PTM data using ReportSites will be a useful addition to existing analysis methods for proteomic studies that produce large data sets. The code used to create ReportSites is freely available from the authors and is hosted online (http://ptmtools.portjackson.org/) for public use.

Authors' Contributions

A.V.G.E conceived the study, carried out phospho peptide enrichments, mass spectrometry and initial data analysis, drafted the manuscript and provided proteomic guidance of coding and interpretation of data. G.J.E wrote code for ReportSites. MRL assisted in mass spectrometry and enrichments. SJC assisted in proteomic guidance of coding, data interpretation and drafted the manuscript. All authors read and approved the final version of the manuscript.

Supplementary Information

Supplementary Table 1 – Description of ReportSites code functionalities.

Acknowledgements

A.V.G.E was funded by an Australian Postgraduate Award and a project grant from The Lundbeck Foundation. This work was funded, in part, by the National Health and Medical Research Council of Australia (NHMRC 571002 to S.J.C.) and The Lund beck Foundation (Junior Group Leader Fellowship to M.R.L, project grant to A.V.G.E.).

References

1. Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, et al. (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci Signal 3: ra3.
2. Oberprieler NG, Lemeer S, Kalland ME, Torgersen KM, Heck AJ, et al. (2010) High-resolution mapping of prostaglandin E2-dependent signaling networks identifies a constitutively active PKA signaling node in CD8+CD45RO+ T cells. Blood 116: 2253-2265.
3. Schwartz D, Gygi SP (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat Biotechnol 23: 1391-1398.
4. Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jørgensen TJ (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol Cell Proteomics 4: 873-886.
5. Rudrabhatla P, Grant P, Jaffe H, Strong MJ, Pant HC (2010) Quantitative phosphoproteomic analysis of neuronal intermediate filament proteins (NF-M/H) in Alzheimer's disease by iTRAQ. FASEB J 24: 4396-4407.
6. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105-132.
7. Thingholm TE, Jensen ON, Larsen MR (2009) Analytical strategies for phosphoproteomics. Proteomics 9: 1451-1468.
8. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24: 1285-1292.
9. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28: 3442-3444.
10. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, et al. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res 36: D695-D699.
11. Peri S, Steen H,Pandey A (2001) GPMAW--a software tool for analyzing proteins and peptides. Trends Biochem Sci 26: 687-689.
12. Zwang NA, Hoffert JD, Pisitkun T, Moeller HB, Fenton RA, et al. (2009) Identification of phosphorylation-dependent binding partners of aquaporin-2 using protein mass spectrometry. J Proteome Res 8: 1540-1554.