Journal of
Proteomics & Bioinformatics

Edwards et al., J Proteomics Bioinform 2014, 7:12
DOI: 10.4172/jpb.1000344

Research Article Open Access

A Tool for Mapping and Displaying Spatial Patterns and Relationships between Post-Translational Modifications in Two and Three Dimensions

Alistair V G Edwards[1]*, Gregory J Edwards[2] and Martin R Larsen[1]

[1]*Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, Odense, Denmark*
[2]*Port Jackson Bioinformatics, Sydney, Australia*

Abstract

Background: We have previously published a software tool for the analysis of the environments around protein post-translational modification on a sequence-driven basis. This article describes advances to this tool, primarily focused on methods of describing the spatial relationships between post-translational modifications on a 2D and 3D basis.

Results: We show that we are able to extract meaningful information about protein modification distributions and inter-relationships using this software. We show examples of modification clustering relationships based on our own data, and discuss the potential use of this and similar tools in the future.

Conclusions: We anticipate that this tool will be useful in describing how and when post-translational modifications are deployed by cells, thereby revealing novel aspects of cellular regulation.

Keywords: Protein modification; PTM; PTM coding; Three dimension; Software

Introduction

Large-scale proteomic datasets containing large amounts of post-translational modification (PTM) information are now commonplace. 20,000 unique reported sites of phosphorylation, for instance, are not uncommon in a single experiment [1], and with similar increases in data density repeated across many different categories of PTM [1-3]. This increase in available data naturally expands our understanding of how and when proteins are modified by specific types of PTM, though the increase in functional and biological understanding gained is limited by our ability to fully characterize individual sites and/or proteins. These large datasets are, however, also a good resource in other respects. For example, there is increasing attention being given to the biological information that may be contained within the patterns of PTMs on larger scales than single sites and single proteins. Observations of the likelihood for phosphorylation to occur in disordered regions of proteins [4,5], or glycosylation to be present on cell membrane-exposed proteins, for example, have shown that it is indeed possible to draw conclusions about the general usage of PTMs in cells from analysis of PTM locations in large scale datasets.

There is, however, another aspect of the environment around PTM sites that may also have an impact on their distributions: namely, the presence or absence of other PTMs. Many studies in recent years have shown that there exist functional interactions between PTMs; this is known as PTM crosstalk [6-9], and is thought to act to regulate protein behaviour. Furthermore, it has been shown that PTMs often occur in hotspots or clusters of dense modification [2,10]. These positional relationships are conserved over evolutionary time [11], suggesting that they have a functional role. Proposed functions include that the clusters may act as an information integrating mechanism, allowing different combinations and permutations of PTMs to perform different roles (essentially being interpreted as a single unit), thereby increasing the overall complexity and capacity for control of the system [12,13].

Of course, not all PTMs exist in this type of cluster, but knowing which do and how these clusters are altered over time and on a scale of tens of thousands of PTM sites, may well reveal novel aspects of cellular function. Our laboratory has previously created a software tool that has been used to perform assessments of physical environments around PTM sites [14] as well as their proximity to each other on a 2D basis (i.e. separation in terms of number of amino acids) in large datasets, and revealed that PTMs of several categories display significant clustering behaviour in sequence space, and that this behaviour is altered over time [15,16]. Proteins are, however, three dimensional structures and therefore a thorough assessment of PTM clustering tendencies requires an approach that works in three dimensions.

We have therefore extended our software tool to allow users to describe 3D as well as 2D relationships in PTM datasets, by mapping where in the appropriate proteome a PTM site of a given category has been identified, describing how this position relates to all other PTM sites in the proteome (i.e. which sites are generally found close to which other sites, and so on). In our data, generated from samples of developing mouse brains, this has shown that there exists a significant effect in three dimensions for PTMs to be closer than expected by chance.

Materials and Methods

In order to develop and test this code, a large scale PTMomic dataset was produced. Mouse brain developmental samples (0, 8, 21 and 80 days old) were enriched for phosphopeptides using titanium dioxide essentially following the method of Larsen et al. [17] and analysed on an LTQ-Orbitrap Velos (Thermo Scientific, San Jose, CA)

***Corresponding author:** Alistair VG Edwards, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, Odense, Denmark, Tel: +45 6550 2478; E-mail: aled@bmb.sdu.dk

Received November 24, 2014; **Accepted** December 19, 2014; **Published** December 23, 2014

Citation: Edwards AVG, Edwards GJ, Larsen MR (2014) A Tool for Mapping and Displaying Spatial Patterns and Relationships between Post-Translational Modifications in Two and Three Dimensions. J Proteomics Bioinform 7: 385-388. doi:10.4172/jpb.1000344

using higher energy collision dissociation. Raw files were processed and searched using Proteome Discoverer (version 1.0; Thermo Scientific) and Mascot (version 1.12; www.matrixscience.com/) against a Swiss-prot rodent database (UniProtKB/Swiss-Prot 2012_10, 538,259 target sequences). Peptide hits with a low confidence after false discovery rate analysis were discarded.

ReportSites itself is a command-line oriented Perl program of ~9000 lines. 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 doing this, the program also retains information about the relative position of each unique PTM site and any quantitative information regarding the peptide and/or site. Based on this it is possible to centralize each PTM and describe the distributions of other PTMs around it. When this is repeated over many tens of thousands of sites, maps of relative position of PTMs based on sequence are produced (see [15,16] for usage examples). The 2D tool was extended by mapping the SwissProt proteins to PDB (The Protein Databank [18]) protein structure files via the mappings at http://www.ebi.ac.uk/pdbe/docs/sifts/quick.html, in particular uniprot_pdb.csv.gz. Approximately 11.1% of the SwissProt proteins with PTMs were mappable. This is astandard yield of proteins with known 3D structures. Next the sequences with PTMs were cross-mapped from SwissProt to PDB using exact and fuzzy matching of peptide sequences into proteins and chains within the PDB data. The final yield of mappings of experimental PTMs to 3D locations was 13.4%. (It should be noted that a large dataset with a large database will take some time to complete–for reference, our dataset of some 14,000 PTM sites requires in the region of one hour to run.)

The alpha carbon of the residues in the PDB chain were chosen as the "location" of the PTM. The X,Y,Z coordinates of these were accumulated into a suitable data structure for comparing any desired subset of 3D distances between categories of PTMs. 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 if required. 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 tool itself can be downloaded from http://reportsites3d.s3.amazonaws.com/list.html.

Results

ReportSites offers the choice of looking at PTM relationships on a 2D or a 3D level. In the 2D situation, a PTM category, say acetylation, is selected and the frequency of occurrence of other PTMs around all instances of acetylation is summed in bins of one amino acid, over a user-defined range. This results in a visual output of PTM occurrence in the vicinity of acetylation sites (in our example) across an entire dataset. This can of course be applied to any other category of PTM, either as the central site or as one of the PTMs whose frequency should be measured. Output from our own mouse brain dataset in two dimensions is shown in Figure 1, illustrating how the tool allows all PTMs to be displayed relative to each other, allowing a rapid assessment of the propensity for PTM clustering.

On a 3D level, we are now able to ask the same question. That is, if we use physical separation in space instead of separation by a certain number of amino acids as a measure of PTM distributions, will we see a similar clustering pattern? This is an important point, as many PTM sites may be closely spaced in 3D space without being close in sequence space. By accessing 3D crystal structures in an automated

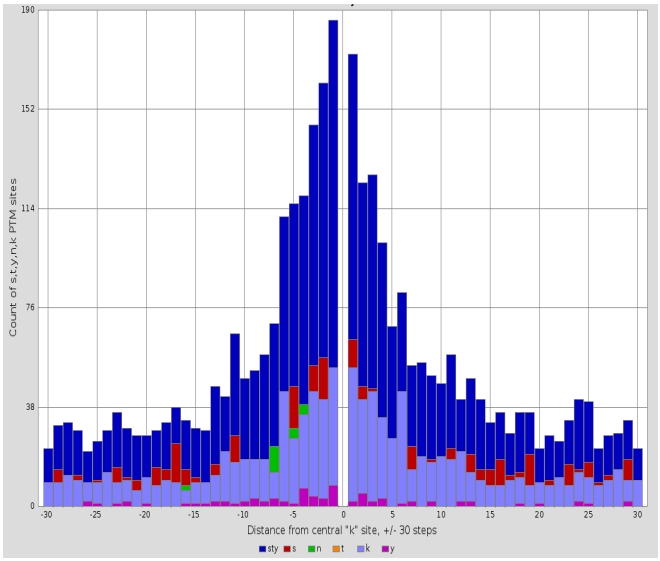

Figure 1: Example of raw 2D PTM distributions around acetylation sites. Frequencies on the Y-axis, number of amino acids away from the centralised PTM category on the X-axis. A clear peak can be seen in the frequencies of other PTMs in the near vicinity of acetylation sites (blue–all phosphorylation; red–phosphoserine; purple–phosphotyrosine; lilac–acetyllysine; green–deamidated asparagine).

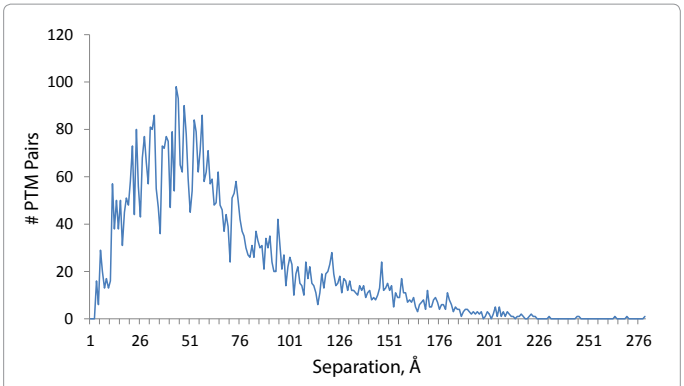

Figure 2: Raw distances in Å between all pairs of PTMs in development dataset. Number of PTM pairs on the Y-axis, distance between members of PTM pairs on the X-axis. A trend can be seen for PTMs to occur at a separation of approximately 25 to 50 Å, though this data is uncorrected.

manner using the PDB we can describe the distance between all PTM pairs in a dataset, and thereby describe the patterns of modification in 3D space and the relationships between PTM pairs. As is the case for 2D analyses, the 3D analysis can be divided by whichever PTM pair is required. The initial analysis of all PTM pairs in mouse brain data is shown in Figure 2.

Two things should be noted with regard to the data in Figure 2: firstly, that it includes PTM pairs which we already know are close in 2D space, and secondly that some pairs of PTMs are evidently relatively widely spaced (>200 Å). Naturally, if a site contributes to a clustering effect in two dimensions, it will also do so in three dimensions, though this does not add to the understanding of the overall clustering tendency. Therefore, we next included a simple, user-defined filter to exclude sites from the 3D analysis which are already close in sequence.

At the same time, we wished to normalize the 3D data to bring sites

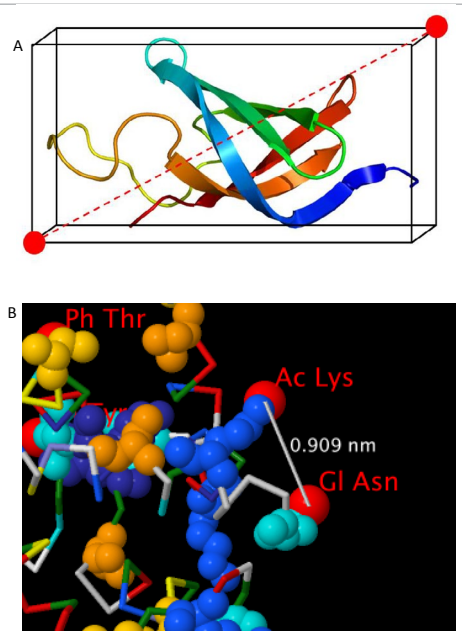

Figure 3: Protein size estimation method. A) Protein size can be effectively estimated using a bounding box method, whereby a 3D box is created around around the protein and measuring its longest axis. **B)** Illustrative output from PTM mapping in 3 dimensions. PTMs marked in red, distances marked in white. All pairs of PTMs can be interrogated.

which may be very widely spaced (e.g. at opposite ends of elongated proteins) into an easily assessable frame. This was done by normalizing the distance in Å between each pair of PTM sites by the size of the protein bounding box-the smallest sized 3D box that could contain the whole protein (see Figure 3A). When these modifications were incorporated into ReportSites, they produced an output as shown in Figure 4A, in this example filtered to remove PTMs closer to each other than 10 residues in sequence. This value was chosen as the majority of the 2D clustering pattern was observed at distances less than 10 residues. Figure 3B shows a conceptual visualisation of the 3D mapping investigations, showing the distance between two chosen PTMs. This visualisation is performed in Jmo [19].

These results show a clear peak in inter-PTM distance, normalized by protein size. The median inter-PTM distance lies in the 27th percentile, which is more than 2 standard deviations from the expected median for a normal distribution of PTMs. Interestingly, despite the presence of large distances in the dataset the upper ranges of the plot (from approximately percentile 70 to 100) are almost empty of data. Additionally, the exclusion of PTM pairs closer to each other than 10 residues had little effect on the distribution.

We assume that a 'normal distribution' in this context has a median of 50. It is possible, however, that a random selection of PTM sites within a given protein would have an average inter-PTM distance which does not reflect the idealised normal distribution. In order to determine if this was the case, we generated a dataset of random points within an idealised cuboid protein, measured the distance between them and normalized them. This produced a dataset as shown in Figure 4B, with a median inter-PTM distance in the 37th percentile, illustrating two things: firstly, that a simple distribution model is inadequate to describe inter-PTM distances, and secondly, that the median distance in our data varies still from the random distribution to a substantial degree.

Discussion

We report an extension of the software tool ReportSites that can facilitate site-specific analysis of large-scale PTMomic data. This tool allows quick and easy assessment of PTM spatial relationships on both a 2D and 3D level. This is of importance as 3D assessment will be necessary to give a 'real-world' picture of the PTM environment in large-scale datasets. This analysis showed a trend for PTM pairs to be closer in three dimensions than would be expected by chance. By this we mean that a random distribution of non-modified residues should produce a curve with a median separation at the 50th percentile, while in our PTM data we see the median at approximately the 26th percentile.

It is interesting that we were able to recapitulate in three dimensions the 2D clustering effects observed in our own work [15,16] as well as that of others [10,11] given that this effect was observed even when closely spaced sites on a sequence basis were excluded from analysis. This suggests that there is a distinct effect functioning in three dimensions, in addition to that already shown in two dimensions. We also observed the same 3D patterns when using data downloaded from the Phosida website [20], which indicates that the effect is not limited to mice or to brains.

The implications of these findings could be interesting–as has been noted elsewhere, patterns of PTM proximity is a prerequisite for certain proposed mechanisms for complex cellular control to exist. These mechanisms primarily focus on an effect called PTM coding, whereby densely modified areas of proteins behave as single units and are read as a 'code' [8,12,13,21,22].

When transitioning from 2D analysis to 3D, a significant drop in data density was observed, due to the lack of 3D structures for many proteins identified by high-throughput approaches. This limits

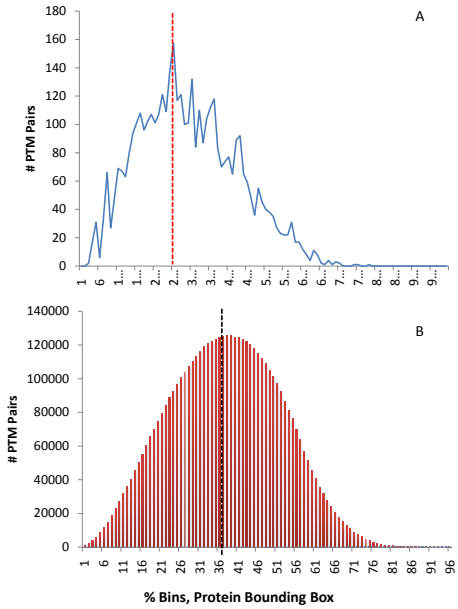

Figure 4: Normalized 3D PTM occurrence. Number of PTM pairs on Y-axis, distance between pairs in percentile protein size on X-axis. A) Distances between PTMs in our experimental datsaset. When corrected for outliers created by large proteins and adjusted to removed sites already close in sequence, a peak in inter-PTM distance was observed that was significantly lower than would be expected by chance. Median value in red, at 27th percentile; B) Equivalent normalized distribution for random PTMs within an idealised cuboid 'protein'. Median value in black, at 37th percentile.

the ability of the tool to use the full extent of the data available to it, but in the absence of other methods to predict protein structure accurately this may be unavoidable. Given the increase in interest in the possibility that PTMs may be used in a combinatorial coding manner, we anticipate that tools such as this will be useful in describing to what degree PTMs are found in clusters of dense modification. This will reveal areas of the proteome as well as individual proteins which attract intense modification, which may well point out key regulatory nodes for the cell. The difference in median inter-PTM distance between our experimental dataset and a set of random PTMs in a regular cuboid 'protein' suggests that we are able to access some degree of genuine biological information with this tool, as opposed to simply sampling the 'noise' in the system.

We note that localisation of PTM sites is an issue in this area. As ReportSites does not contain any localisation aspects itself, the data entered must be filtered to an adequate level of confidence before entry into the program. This includes false-positive assessment as well as e.g. phosphorylation localisation.

We plan on developing this tool further to incorporate measures of protein structure that can be assessed on a high throughput basis. This includes, for example, computational tools such as those to predict protein disorder [23]. At the time of writing the authors are not aware of any other freely available software tools that perform the function of ReportSites with which it can be compared.

Conclusions

Development of ReportSites allowed us to describe more clearly the relationships between PTMs in large datasets, moving from a purely 2D basis to a 3D. This analysis revealed clustering effects in three dimensions which are separate to those observed in two dimensions. When working with a large enough PTMomic dataset (now commonly produced), this tool may aid in revealing novel aspects of the biological functions and cellular uses of post-translational modifications.

Author's Contributions

A.V.G.E conceived the study, carried out all laboratory work 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. All authors read and approved the final version of the manuscript.

Acknowledgement

AVGE was supported by a project grant in biomedicine from the Lundbeck Foundation. Development and online computing services were supported by an Amazon Web Services in Education grant to AVGE. MRL is the recipient of a Lundbeck Foundation Junior Group Leader Fellowship.

Availability

ReportSites is freely available from the authors on request.

References

1. Mertins P, Qiao JW, Patel J, Udeshi ND, Clauser KR, et al. (2013) Integrated proteomic analysis of post-translational modifications by serial enrichment Nat Methods 10: 634-637.
2. Swaney DL, Beltrao P, Starita L, Guo A, Rush J, et al. (2013) Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. Nat Methods 10: 676-682.
3. Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, et al. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science 325: 834-840.
4. Collins MO, Yu L, Campuzano I, Grant SG,Choudhary JS (2008) Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. Molecular and Cellular Proteomics 7: 1331-1348.
5. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, et al. (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32: 1037-1049.
6. Parker BL, Shepherd NE, Trefely S, Hoffman NJ, White MY, et al. (2014) Structural basis for phosphorylation and lysine acetylation cross-talk in a kinase motif associated with myocardial ischemia and cardioprotection. J Biol Chem 289: 25890-25906.
7. Butkinaree C, Park K, Hart GW (2010) O-linked beta-N-acetylglucosamine (O-GlcNAc): Extensive crosstalk with phosphorylation to regulate signaling and transcription in response to nutrients and stress. Biochim Biophys Acta 1800: 96-106.
8. Hunter T (2007) The age of crosstalk: phosphorylation, ubiquitination, and beyond. Molecular Cell 28: 730-738.
9. Kowenz-Leutz E, Pless O, Dittmar G, Knoblich M, Leutz A (2010) Crosstalk between C/EBP[beta] phosphorylation, arginine methylation, and SWI/SNF/Mediator implies an indexing transcription factor code. Embo J 29: 1105-1115.
10. Beltrao P, Albanese V, Kenner LR, Swaney DL, Burlingame A, et al. (2012) Systematic functional prioritization of protein posttranslational modifications. Cell 150: 413-425.
11. Minguez P, Parca L, Diella F, Mende DR, Kumar R, et al. (2012) Deciphering a global network of functionally associated post-translational modifications. Mol Syst Biol 8: 599.
12. Creixell P, Linding R (2012) Cells, shared memory and breaking the PTM code. Mol Syst Biol 8: 598.
13. Nussinov RT, sai CJ, Xin F, Radivojac P (2012) Allosteric post-translational modification codes. Trends Biochem Sci 37: 447-55.
14. Edwards AVG, Edwards G, Larsen MR, Cordwell SJ (2012) ReportSites - a computational method to extract positional and physico-chemical information from large-scale proteomic post-translational modification datasets. Journal of Proteomics and Bioinformatics 5: 104-107.
15. Edwards AV, Edwards GJ, Schwammle V, Saxtorph H, Larsen MR (2014) Spatial and temporal effects in protein post-translational modification distributions in the developing mouse brain. J Proteome Res 13: 260-267.
16. Edwards AV, Schwammle V, Larsen MR (2014) Neuronal process structure and growth proteins are targets of heavy PTM regulation during brain development. J Proteomics 101: 77-87.
17. Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jorgensen TJD (2005) Highly Selective Enrichment of Phosphorylated Peptides from Peptide Mixtures Using Titanium Dioxide Microcolumns. Molecular and Cellular Proteomics 4: 873-886.
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
19. Herraez A (2006) Biomolecules in the computer: Jmol to the rescue. Biochem Mol Biol Educ 34: 255-261.
20. Gnad F, Gunawardena J, Mann M (2010) PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res 39: D253-D260.
21. Minguez P, Letunic I, Parca L, Bork P (2013) PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. Nucleic Acids Res 41: D306-311.
22. Lothrop AP, Torres MP, Fuchs SM (2013) Deciphering post-translational modification codes. FEBS Lett 587: 1247-1257.
23. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, et al. (2003) Protein disorder prediction: implications for structural proteomics. Structure 11: 1453-1459.